TITULNÍ LIST PERIODICKÉ ZPRÁVY 2008 PROJEKTU 2C06009

Ministerstvo školství, mládeže a tělovýchovy

2C06009

PROSTŘEDKY TVORBY KOMPLEXNÍ BÁZE ZNALOSTÍ PRO KOMUNIKACI SE SÉMANTICKÝM WEBEM V PŘIROZENÉM JAZYCE

řešitel - **doc. Ing. Karel Ježek, CSc.**

.....

(podpis)

za příjemce - koordinátor - **Západočeská univerzita v Plzni** (IČ: 49777513)

rektor

Doc. Ing. Josef Průša, CSc.

.....

(podpis, razítko)

Verze zprávy: **1** Zpracováno dne: **29.1.2009**

2. SKUTEČNOST ZA UPLYNULÉ OBDOBÍ - 2008

2.1. PROJEKTOVÝ TÝM A ŘEŠITELSKÉ TÝMY

2.1.1. PROJEKTOVÝ TÝM

IČ organizace	49777513
Obchodní jméno - název	**Západočeská univerzita v Plzni**
Zkratka názvu	ZČU
Role organizace	příjemce - koordinátor
Vazba na organizaci	00216224
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Univerzitní 8/
- PSČ, obec	30614 Plzeň
- stát	Česká republika
- telefon	377 631 111
- http://	www.zcu.cz
Bankovní spojení	
-DIČ	CZ49777513
- banka kód, název	0100 - Komerční banka, a.s., Plzeň
- číslo účtu, sp.symbol	4811530257,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Doc. Ing. Josef Průša CSc.
- funkce	rektor
- telefon	377631000
- mobil	606665105
- fax	377631002
- email	rektor@rek.zcu.cz

IČ organizace	00216224
Obchodní jméno - název	**Masarykova univerzita**
Zkratka názvu	MU
Role organizace	spolupříjemce
Vazba na organizaci	49777513
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Žerotínovo náměstí 617/ 9
- PSČ, obec	60177 Brno
- stát	Česká republika
- telefon	549 491 1111
- http://	www.muni.cz
Bankovní spojení	
-DIČ	CZ00216224
- banka kód, název	0100 - Komerční banka Brno-město
- číslo účtu, sp.symbol	85636621,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Prof. PhDr Petr Fiala PhD
- funkce	rektor
- telefon	549491001
- mobil	
- fax	
- email	rektor@muni.cz

2.1.2. ŘEŠITELSKÝ TÝM

Celé jméno, RČ	**Albrecht Štěpán Ing.** 810520/2061 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 496 377 632 402 albrs@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	20

Celé jméno, RČ	**Bártek Luděk Mgr.** 7201083791 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3215 bar@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Brada Přemysl Ing. PhD. MSc.** 7007012111 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	3772435 brada@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Češka Zdeněk Ing.** 8207311244 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 zceska@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	31.25

Celé jméno, RČ	**Ekštein Kamil Ing. PhD.** 7705302011 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 kekstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Habernal Ivan Ing.** 830705/1764 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 habernal@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	10

Celé jméno, RČ	**Hanks Patrick Ph.D.** 400324 GB
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	hanks@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	

Celé jméno, RČ	**Hejtmánek Jan Ing.** 821101/2095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 hejtman2@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	10

Celé jméno, RČ	**Horák Aleš PhD.** 7409014250 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 4377 hales@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Hynek Jiří ing. PhD.** 720506/2029 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632455 hynekj@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Ježek Karel doc. Ing. CSc.** 420617110 CZ
Role osoby při řešení projektu	řešitel
Spojení	377 632 475 jezek_ka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Klečková Jana doc. Dr. Ing.** 496108095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 421 kleckova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Konopík Miloslav Ing.** 8103261782 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 konopik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Kopeček Ivan doc. RNDr. CSc.** 490303075 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3861 kopecek@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Král Pavel Ing. PhD.** 760317/2049 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632454 pkral@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Krutišová Jana Ing.** 5955160046 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 413 krutisova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Lobaz Petr Ing.** 7607292033 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 447 lobaz@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky a výpočetní techniky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	5

Celé jméno, RČ	**Matoušek Václav prof. Ing. CSc.** 480613108 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 471 matousek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mautner Pavel Ing. PhD.** 6505222592 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 mautner@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mouček Roman Ing. PhD.** 7607072000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 moucek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pala Karel doc. PhDr. CSc.** 390615416 CZ
Role osoby při řešení projektu	spoluřešitel
Spojení	549 49 5616 pala@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Pavelka Tomáš Ing.** 7909182083 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Pomikálek Jan Mgr.** 7910090419 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 1864 xpomikal@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	

Celé jméno, RČ	**Ptáčková Helena** 705914/2079 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 463 377 632 402 ptackova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Rohlík Ondřej Ing. PhD.** 7510031925 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632450 rohlik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Rychlý Pavel Mgr. PhD.** 7301235359 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 6399 pary@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Sojka Petr RNDr. PhD.** 6309171000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549496966 sojka@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Steinberger Josef Ing. PhD.** 7909182127 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 479 jstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Tesař Roman Ing.** 7909302379 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 romant@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	10

Celé jméno, RČ	**Toman Michal Ing.** 8007042054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 mtoman@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Zíma Martin Ing. PhD.** 7405042073 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632431 zima@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Fakulta aplikovaných věd Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

2.1.3. ZMĚNY V PROJEKTOVÉM A ŘEŠITELSKÝCH TÝMECH - rok 2008

Pč.	Typ	Popis
1	změny v projektovém týmu a řešitelských týmech	Ing. Štěpán Albrecht ukončil činnost v řešitelském týmu k 30. dubnu 2008 z důvodu odjezdu na zahraniční stáž.
2	změny v projektovém týmu a řešitelských týmech	K 1.5.2008 nastoupili do řešitelského týmu doktorandi Ing. Ivan Habernal a Ing. Jan Hejtmánek, kteří byli přijati na katedru jako interní doktorandi od 1.9.2007.
3	změny v projektovém týmu a řešitelských týmech	V únoru dokončil disertaci a plánované aktivity a odešel do průmyslové praxe ing. Roman Tesař, PhD.
4	změny v projektovém týmu a řešitelských týmech	V listopadu byl do řešitelského týmu zařazen ing. Petr Lobaz.

2.2. ČASOVÝ POSTUP PRACÍ

Komentář k metodice a časovému postupu prací a průběhu aktivit za uplynulé období

Metodika a časový rozvrh postupu prací byly dodrženy. Během řešení se však vyskytla potřeba doplnění několika dalších dílčích úkolů, neboť některé dílčí činnosti, jejichž provedení bylo původně plánovano v rámci jiných činností, se ukázaly natolik rozsáhlé, že byly do plánu řešitelských prací a také do předkládané zprávy doplněny jako samostatné. Jinak se plán řešitelských prací nezměnil a byl plně dodržen.

2.2.0. PŘEHLED DÍLČÍCH CÍLŮ SCHVÁLENÉ- SKUTEČNOST 2008

	Číslo	Dílčí cíl podrobně	Datum plnění
	1	**Dílčí cíl** Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím. **Indikátory dosažení - výsledky dílčího cíle** a) Vytvoření uživatelského rozhraní pro hlasový vstup / příp. výstup, které bude použito pro komunikaci se sémantickým webem, a pro jeho podporu vytvoření robustního ASR systému pro inflexní jazyky. K tomu bude nutno vytvořit kvalitní korpus pro ASR a z něj extrahovat dostatečné množství trénovacích dat. V jednotlivých etapách bude v průběhu let 2006 – 2007 vytvořen: - kvalitní audio-korpus pro natrénování systému ASR, - korpus pro natrénování jazykových modelů. b) Příprava datových kolekcí a pomocných rutin vyhledávacího systému ve vícejazyčných korpusech, včetně prostředků pro zpřesňování uživatelských dotazů pomocí thesauru a nástrojů pro disambiguaci víceznačných slov, na bázi klient/server aplikace. Jednotlivé dílčí výsledky řešení projektu lze charakterizovat takto: - vytvoření multijazykových korpusů – základní výběr zahrnuje angličtinu a češtinu, dle možností alespoň některé úlohy plánujeme provádět i se slovenštinou (zajímavá je blízkostí k češtině) a němčinou, - metoda automatického rozpoznání jazyka – kombinace „stop slov“ a frekvenčních znakových metod. c) Příprava datových kolekcí a modulů pro filtraci a sumarizaci textů: - vytvoření sumarizačních korpusů (pro angličtinu plánujeme využít standardních korpusů, např. DUC a pro češtinu bude vytvořen vlastní, složený vesměs z textů novinových článků, - sumarizace textů založená na latentní sémantické analýze (LSA), vytvoření anotované kolekce pro sumarizátor založený na LSA - vytvoření vícejazyčných korpusů, - rozšíření standardních textových korpusů o korpusy závadných dokumentů pokrývající problematická témata definovaná v zadání. d) Korpus syntaktických stromů (treebank): - korpus bude morfologicky označkován a zjednoznačněn, - bude v něm vyznačena závislostní struktura věty i jednotlivé větné složky včetně koreferenčními vztahy, - korpus bude z části založen na existujícím PDT. e) Korpus vzorových přepisů vybraných vět a jejich sémantické reprezentace: - text korpusu bude podmnožinou korpusu syntaktických stromů, - ve stromech budou vyznačeny významy z dostupných ontologií (WordNet), - věty budou rozšířeny o logické formy. f) Doplnění morfologického značkovače o robustní hádací proceduru, která bude spolehlivě přiřazovat morfologické značky i neznámým slovům. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Jedná se o vytvoření podpůrného aparátu, bez něhož nelze další zamýšlené cíle projektu dosáhnout. Vytvořeny budou proto korpusy v podobě rozsáhlých datových souborů se specifickou strukturou a organizací a pro jejich údržbu a prohledávání budou vyvinuty speciální softwarové nástroje. Výsledky budou soustředěny do soustavy datových souborů a její obsah prezentován formou publikace na konferencích a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující náplň dílčího cíle „1“ a nástin jejich řešení jsou následující: RF1: Během zpracování korpusů a korpusových nástrojů se vyskytnou další korpusy obsahující srovnatelná data. Řešení: Korpusy pro český jazyk vznikají v ČR na celkem pěti pracovištích, která udržují těsné kontakty a výsledky výzkumu si vzájemně vyměňují nebo se o nich poměrně obsáhle informují. Navíc je třeba rozlišovat mezi korpusy psanými (textovými) a řečovými. Řečové korpusy vznikají prakticky jen na pracovištích v Plzni, Brně a Liberci, z nichž dvě se na řešení	- 31.12.2007

		tohoto projektu budou podílet. Navíc vznik jakéhokoli dalšího korpusu je pozitivním jevem, neboť v tomto oboru více než kdekoli jinde platí, že vhodných dat není nikdy dostatek. Tudíž korpusy vytvořené v rámci navrhovaného projektu budou v každém případě využity i dalšími pracovišti. V případě cizojazyčných korpusů budou využívány korpusy, které jsou k dispozici v systému ELRA (European Language Resources Association). RF2: Nepodaří se získat dostatek materiálů, resp. mluvčích, pro vytvoření textových, resp. audiokorpusů. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť již současný web poskytuje doslova nepřeberné množství textového materiálu, z nichž lze za použití vhodných vyhledávacích metod vybrat dostatečné množství materiálu pro vytvoření korpusu. V případě řečových korpusů nejde ani tak o problém nalezení vhodné množiny dat nebo množiny vhodných mluvčích, nýbrž kritickým faktorem je čas. Pořizování řečových dat a zejména jejich následné zpracování (třídění, anotace, apod.) vyžaduje značné množství času, avšak riziko lze úspěšně odstranit kvalitním managementem projektu. RF3: V průběhu naplňování dílčího cíle projektu se vyskytne komerční software řešící problematiku pořizování korpusů. Řešení: Pokud se nějaký software vyskytne a bude využitelný, nebude díky modularitě předpokládaného programového vybavení příliš obtížné ho do vytvářeného software začlenit. Pravděpodobnost jeho výskytu v dohledné době je však minimální.	
	2	**Dílčí cíl** Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka. **Indikátory dosažení - výsledky dílčího cíle** a) Návrh formalismu pro popis sémantiky na rozsáhlejší doméně, návrh vhodně strukturovaného sémantického popisu dotazů uživatelů, eventuálně vytvoření vlastního hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému. b) Vytvoření ontologií pro aplikaci formalismu popisujícího sémantiku. Jednotlivými výsledky budou: - návrh ontologie, sémantických konceptů – datový formát XML, vytvoření UML modelu, - návrh ohodnocení jednotlivých konceptů vektorem sémantických příznaků, a to jak doménových, tak obecnějšího charakteru, - návrh soustavy vektorů ohodnocení jednotlivých konceptů. c) Vytvoření multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí, jeho zakomponování do prostředí pro vyhledávání a vývoj metod ohodnocování jeho kvality, návrh metod disambiguace v multijazykovém prostředí s využitím kontextu, thesauru a pravděpodobnostních metod: - sumarizační systém obohacený o kompresi souvětí, - systém rezoluce anafor a jeho využití při sumarizaci – pro angličtinu bude využit systém GuiTAR, vytvořený na univerzitě Essex (Anglie), pro češtinu bude na základě poznatků získaných na českých pracovištích vytvořen vlastní systém, - metoda hodnocení kvality sumarizátorů na základě LSA. d) Vývoj nových, dokonalejších modelů elektronických dokumentů tak, aby při použití textových klasifikačních algoritmů bylo dosaženo co nejlepších výsledků při rozpoznávání tématu, rozpoznávání spamových emailů, detekci dokumentů se závadným obsahem apod. e) Vytvoření metodologie a nástrojů pro analýzu webových dokumentů. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Při naplňování tohoto dílčího cíle půjde o vytvoření základního teoretického podpůrného aparátu, bez něhož nebude možné další kroky realizovat. Jediný tento dílčí cíl bude mít charakter spíše základního výzkumu – půjde o vývoj metod, metodologií a formálních modelů pro návrh zamýšleného komunikačního rozhraní, avšak součástí výzkumných prací bude též experimentální implementace a vytvoření softwarových nástrojů pro evaluaci vyvíjených metod a formalismů. Výsledky budou shrnuty do písemných dokumentů a prezentovány téměř výhradně formou publikací na konferencích, v odborných časopisech a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „2“ a nástin jejich řešení mohou být následující: RF1: Nepotvrzení či neplatnost výzkumných hypotéz poskytujících základ pro vytvoření	- 31.12.2008

		formalismů a modelů. Řešení: Plánovaný dílčí cíl zde nestojí na jediné výzkumné hypotéze, nýbrž na teoretickém základu návrhu komunikačních systémů. Využito bude jak dosavadních poznatků z návrhu existujících komunikačních rozhraní a systémů pro interakci člověka s počítačem, tak i poznatků z psychologie komunikace a doporučení TC.13 IFIP (for HCI). Základním rizikem proto bude opět časový faktor, který lze výrazně omezit dobrým managementem projektu. RF2: Nedostatečná erudice členů týmu pro vývoj formálních prostředků. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť oba participující týmy jsou složeny minimálně z poloviny ze starších zkušených výzkumníků, z nichž někteří se předmětnou oblastí zabývají 25 i více let, z druhé části pak z mladých perspektivních pracovníků, kteří buď vyrostli anebo se podíleli na řešení podobné problematiky a potřebné teoretické základy oboru již získali, zejména v doktorandském studiu.	
	3	**Dílčí cíl** Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. **Indikátory dosažení - výsledky dílčího cíle** a) Implementace uživatelského rozhraní pro hlasovou komunikaci se sémantickým webem –součástí výsledku budou: - implementace LVCSR rozpoznávače, - natrénování akustických a jazykových modelů, - implementace nahrávacího modulu se stochastickým modelem detekce řečového signálu, - implemetace parametrizátoru na bázi MFCC, - návrh a implementace modulu pro akustické modelování založeného na umělých neuronových sítích nebo směsích Gaussových funkcí, - návrh a implementace efektivního dekódovacího algoritmu, který dokáže pracovat s gramatikami a stochastickými jazykovými modely, - programová realizace a ověření funkčních vlastností robustního ASR systému pro inflexní jazyky. b) Systém pro extrakci významu ze spontánních promluv – dílčími kroky k dosažení tohoto dílčího cíle budou: - návrh a realizace optimální řečové databáze, - návrh systému sémantického značkování řečových dat, - báze znalostí umožňující automatizované či automatické značkování spontánních promluv uložených v databázi, - implementace stochastických sémantických gramatik pro automatickou sémantickou analýzu dotazu uživatele, - využití hierarchické ontologie pro tvorbu strukturalizovaného popisu dotazů uživatele a pro zajištění schopnosti zobecňování z natrénovaných dat, - aplikace metod mělkého (shallow) parsingu promluv pro částečnou analýzu dotazů uživatele. c) Vytvoření komfortního uživatelského rozhraní pro práci se sémantickým webem – součástí tohoto dílčího cíle bude: - návrh příslušného dialogového manageru akceptujícího tzv. kombinovanou iniciativu ve vedení dialogu (mixed initiative), - vytvoření robustního systému pro efektivní a časově nenáročné vyhledávání dat v řečové databázi, - vytvoření robustního a spolehlivého modelu sémantické hierarchie a jeho implementace. d) Aplikace a modifikace OWL standardu v českém prostředí. e) Aplikace klasifikačních metod v multijazykovém prostředí. f) Kompletace multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí. g) Algoritmy vhodné pro generování itemsetů a n-gramů a ověření jejich úspěšnosti pro klasifikaci textových dokumentů. h) Výchozí algoritmy pro vyvozování nových znalostí z informací získaných z volného textu. i) Prototyp programu pro přiřazování logických formulí větám z volného textu. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** V dílčím cíli „3“ jde o vytvoření souboru programových produktů, které vzniknou implementací teoretických metod a formalismů vytvořených v rámci dílčího cíle „2“. Výsledky budou mít jednoznačně aplikační charakter, i když vesměs půjde jen o experimentální software, bez něhož nelze metody a modely verifikovat. Výsledky však bude možno předat i dalším zájemcům, protože se předpokládá úplná dokumentace vytvořeného programového vybavení.	- 31.12.2009

		Výsledky budou prezentovány jako balíky experimentálního software a metod, dále budou publikovány na konferencích, v průběžných výzkumných zprávách, případně také zveřejněny formou speciálních letáků, v tisku a uvažuje se též o možnosti předvedení na specializovaných veletrzích a výstavách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „3“ a nástin jejich řešení: RF1: V průběhu projektu přestane být o vytvářené přístupové technologie zájem a pracoviště účastnící se na řešení projektu se tak ocitnou bez reálné využitelnosti svých výsledků. Řešení: Současným trendem je naopak příklon k využívání multimediálních a multimodálních dat, ukládání velkých množství dat a informací na běžných počítačových prostředcích, sílí propojování informačních technologií s rozhlasovým a televizním vysíláním, streamovanými médii a mobilními komunikacemi. Nové hardwarové prostředky budou vyžadovat nové technologie přístupu k datům, přičemž preferována bude komunikace v přirozeném jazyce, ať už psanou nebo mluvenou formou. Vyvíjené programové prostředky tento trend jednoznačně podpoří a proto je toto riziko za dobu řešení projektu téměř nulové. RF2: V průběhu řešení projektu se vyskytne komerční software řešící problematiku srovnatelnou s předpokládanými výsledky projektu. Řešení: Komerční řešení využívající přístup k datům na webu prostřednictvím přirozeného jazyka jsou dosud v plenkách a komerční sféra naopak aktivně vyhledává zajímavé práce z akademické sféry. Proto je toto riziko minimální, očekáváme naopak velký zájem z komerční sféry. RF3: Časové faktory ovlivňující zpracování software. Řešení: Při implementaci a programové realizaci metod vyvinutých v rámci dílčího cíle „2“ může dojít k určité časové tísni vlivem nevhodně zvolených implementačních nástrojů, eventuálně nezkušeností některých mladších členů týmu. Riziko je však minimální, neboť řešitelský kolektiv je složen vesměs ze zkušených výzkumníků a mladých pracovníků, kteří již obdobné, i když jednodušší systémy v minulosti vytvářeli a implementovali. Časový faktor lze navíc výrazně ovlivnit dobrým managementem projektu.	
	4	**Dílčí cíl** Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí. **Indikátory dosažení - výsledky dílčího cíle** a) Testování a ověřovací provoz implementovaného hlasového rozhraní – součástí bude - otestování zpracovaného LVCSR rozpoznávacího systému, - ověření funkčních vlastností robustního ASR systému na vhodné množině uživatelů, - otestování vyvinutých metod automatické sémantické analýzy dotazů. b) Ověření funkčních vlastností vytvořených ontologií a hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému analýzy sémantiky, c) Ověření vlastností algoritmů pro klasifikaci a analýzu dat na různých typech dokumentů. d) Otestování a ověření navržených metod na konkrétních typových řešeních, např. na přístupu k webovým stránkám výzkumných a vzdělávacích institucí. e) Vyhodnocení úspěšnosti jednotlivých fází analýzy volného textu od morfologické úrovně až po převod do logických formulí. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Náplní dílčího cíle „3“ je provedení rozsáhlých testů (tzv. field experiments) vyvinutých metod, metodologií, modelů a vytvořeného souboru programových produktů. Předpokládá se testování produktů na obvyklých třech skupinách uživatelů – v prvním kroku budou vlastnosti systémů a metod prověřovány úzkou skupinkou řešitelů projektu, ve druhém kroku bude testovací množina uživatelů vytvořena ze spolupracovníků, kteří však s řešením projektu neměli nic společného a o výsledcích řešení jsou jen velmi kuse informováni, a teprve ve třetím kroku bude systém testován libovolnými uživateli, tzv. „lidmi z ulice“. Zčásti však v tomto kroku budou využiti studenti, kteří všeobecně mají tendenci takové systémy „pokořit“. Výsledky budou kompletně dokumentovány a z vyhodnocení experimentů budou vyvozovány příslušné závěry, tj. systém a jeho části budou průběžně doplňovány, upravovány a opětovně testovány. V závěru budou výsledky testování a ověřovacího provozu publikovány v časopisech, na konferencích a obšírně v závěrečné výzkumné zprávě. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „4“ a možná řešení:	- 31.12.2010

RF1: V průběhu testů se projeví nedostatky v koncepci systému vedoucí k závažným problémům ve funkci systému.
Řešení: Řešitelský tým je složen z odborníků, kteří obdobné, i když jednodušší, systémy již vytvořili a mají z jejich tvorby nezanedbatelné zkušenosti. Tým byl dále doplněn o mladé pracovníky, kteří se podíleli na tvorbě řady produktů pro prezentace na webových stránkách a je jim problematika přístupu k webu velmi blízká. Riziko volby nevhodné koncepce je proto minimální.

RF2: V průběhu testů se projeví nedostatky v implementaci systému a metod.
Řešení: Obdobné jako předchozí rizikový faktor – řešitelský tým je složen z odborníků, kteří obdobné, systémy již vytvořili a mají i z jejich implementace poměrně rozsáhlé zkušenosti. Riziko závažných implementačních chyb je proto minimální, drobné nedostatky v implementaci bývají zpravidla v krátké době snadno odstranitelné.

RF3: Nepodaří se vytvořit dostatečně reprezentativní množiny testovacích osob.
Řešení: Ve vztahu k odstavci 3.3.3. (tři úrovně testování) je riziko nedostatečného vytvoření skupin testujících osob nepatrné – obě participující pracoviště jsou poměrně rozsáhlá a množinu osob testujících vlastnosti systému nebude problém vytvořit; ostatně bylo již ověřeno v minulosti na jednodušších úlohách. Otázka volby třetí skupiny osob je spíše otázkou vytvořeného přístupu k systému – zde se nabízejí dvě možnosti: Buď si osoby vhodné k testování systému vybírat podle určitých hledisek (bylo tak někdy postupováno v minulosti a osoby byly k testování zvány na řešitelské pracoviště) nebo zveřejnit přístupový portál systému a dovolit testování systému široké veřejnosti prostřednictvím internetu, popř. přes telefon (telefonní přístup je však v současných podmínkách omezen kvalitou spojení v mobilních sítích, resp. kvalita spojení je dána úrovní signálu v místech, kde se potenciální uživatel právě nachází, a výsledky testů jím mohou být zkresleny). Rizikový faktor může být opět minimalizován vhodnými rozhodnutími, resp. dobrým managementem projektu.

2.2.1. AKTIVITY USKUTEČNĚNÉ v roce 2008

Číslo aktivity

01/08

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Sémantické anotování korpusu a vytváření anotačních schémat - dokončení

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

V rámci této aktivity byla vytvořena anotační schémata pro zbylá témata, která nebyla pokryta v aktivitě 2007-04 (nakupování, restaurace, městské a státní úřady, památky, muzea, vlakové a autobusové spoje). Tato schémata byla poté použita pro dokončení prací na anotacích sémantického korpusu. V rámci této aktivity byl dále zpracováván korpus typických dotazů získaný při zpracování aktivity 2006-04.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem aktivity je vytvoření uceleného, sémanticky anotovaného korpusu a sada anotačních schémat. Anotováno bylo všech 20292 připravených a zaznamenaných vět, mezianotátorská shoda dosáhla 71%.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Anotovaný korpus je uložen v datovém úložišti projektu. Dosaženou mezianotátorskou shodu lze ověřit navrženým a implementovaným algoritmem.

Číslo aktivity

02/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Sémantická analýza lexikálních tříd

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

Cílem této aktivity bylo navázat na aktivitu 2007-06 - identifikace lexikálních tříd. Pro danou množinu generických lexikálních tříd (datum, čas, číslice, atd.) byly vytvořeny sémantické gramatiky používající formalismus tzv. aktivních tagů. Byla vyvinuta knihovna umožňující syntaktickou analýzu lexikálních tříd s aktivními tagy s následným vyhodnocením a extrakcí významu ve strojovém formátu.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem aktivity je algoritmus pro převod lexikálních tříd do strojově zpracovatelného formátu (vyjádření obsahu ve vyšším programovacím jazyce). Kontrola výsledku byla provedena testováním algoritmů na množině vět získaných v rámci aktivity 2008-01 a měřením úspěšnosti správné extrakce sémantiky dané lexikální třídy, která se pro dané lexikální třídy pohybovala v rozmezí 95 - 97% .

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky aktivity byly prezentovány na mezinárodní konferenci Text, Speech and Dialogue 2008 v publikaci:
HABERNAL, I. KONOPÍK, M. - Active Tags for Semantic Analysis, In: Text, Speech and Dialogue, 2008. Springer Verlag, Berlin, 2008, s. 69-76. ISSN 0302-9743, ISBN 978-3-540-87390-7.

Softwarový produkt LINGVOParser, což je sémantický parser s podporou aktivních tagů, je volně ke stažení na stránkách projektu http://liks.fav.zcu.cz/mediawiki/index.php/LINGVOParser

---

Číslo aktivity
03/08
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Vývoj základního algoritmu pro automatickou sémantickou analýzu dotazů
Zahájení aktivity
2.1.2008
Ukončení aktivity
31.12.2008
Popis aktivity
V rámci této aktivity byly vytvořeny tři typy algoritmů pro sémantickou analýzu dotazů. Třetí typ algoritmu dosahuje z výše zmíněných nejlepších výsledků. Pro analýzu používá pravděpodobnostní bezkontextovou gramatiku. Součástí algoritmu je modul pro trénování gramatiky a modul pro analýzu vstupní věty. Pro trénování byl použit anotovaný sémantický korpus vytvořený v rámci aktivit 2007-05 a 2008-02, takže výsledky této aktivity úzce souvisejí s aktivitou 2008-02, protože lexikální třídy tvoří základ pro sémantickou analýzu (je používán hybridní pravidlový a stochastický přístup). Výsledkem aktivity je tzv. baseline (základní) verze algoritmu pro sémantickou analýzu. Algoritmus bude v dalších fázích zpracování projektu dále vylepšován a zdokonalován.
Skutečné Indikátory dosažení - výsledky aktivity
Algoritmus byl testován na části dat, která nebyla použita pro trénovaní algoritmu. Úspěšnost algoritmu byla měřena v procentech shody s ručně vytvořenými sémantickými stromy. Dosažená úspěšnost 62% představuje dobrý základ pro další vývoj algoritmu. Robustnost algoritmu byla ověřena na větách získaných v rámci aktivity 2007-37 a lze konstatovat, že algoritmus pracoval korektně pro všechny vstupní testovací věty.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Navržený algoritmus a dosažené výsledky byly publikovány v odborné literatuře uvedené níže. Funkci algoritmu je možno ověřit stažením programu a testovací kolekce ze webových stránek projektu.

Publikace:
HABERNAL, I. KONOPÍK, M.: Lexical Class Semantic Analysis. In: Proceedings of 9th International PhD workshop on Systems and Control, Slovenia, 2008. ISBN 978-961-264-003-3
KONOPÍK, M. HABERNAL I.: Stochastic Parsing in a Hybrid Semantic Analysis System. In: Proceedings of 9th International PhD workshop on Systems and Control, Slovenia, 2008. ISBN 978-961-264-003-3

Číslo aktivity
04/08
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Trénování akustických modelů
Zahájení aktivity
1.12.2007
Ukončení aktivity
20.12.2008

Popis aktivity

V rámci této aktivity byly trénovány akustické modely založené na neuronových sítích a směsích Gaussových funkcí. Úspěšnost rozpoznávání získaná při trénování novým korpusem spontánní řeči byla poměrně nízká a přidání nových dat ke stávajícím korpusům nevedlo ke zlepšení. Důvody, proč se tak stalo, budou předmětem dalšího zkoumání. Podstatná část výzkumu byla zaměřena na zkoumání možností akustického modelování při použití kontextově závislých fonetických jednotek. Největších úspěchů bylo dosaženo s trifóny, ale byla rovněž zkoumána alternativa fonetické jednotky založené na slabikách. Dosažené úspěšnosti rozpoznávání jsou u slabik nižší než u trifónů. Důvodem může být volba metod svazování parametrů trifónů (svazování parametrů řeší problém nedostatku dat pro málo frekventované fonetické jednotky). Pro trifóny používáme tzv. decision tree clustering, který se na slabiky zatím nepodařilo aplikovat. Výsledky experimentů a návrhy možných vylepšení do budoucna jsou uvedeny v článku [2]. V pracích [1] a [3] jsme se zabývali porovnáním akustických modelů založených na neuronových sítích a směsích Gaussových funkcí. Ukázalo se, že v některých případech může použití neuronových sítí vést k větší rychlosti rozpoznávání.

Skutečné Indikátory dosažení - výsledky aktivity

Natrénované akustické modely pro automatický rozpoznávač řeči JLASER.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Popis experimentů provedených s trénováním akustických modelů je uveden v následujících článcích publikovaných na zahraničních konferencích a v jedné disertační práci:

[1] Pavelka, T., Král, P.: Neural Network Acoustic Model with Decision Tree Clustered Triphones. Proceedings of 2008 IEEE International Workshop on Machine Learning for Signal Processing, Cancun, Mexico, 2008, ISBN 978-1-4244-2376.

[2] Hejtmánek, J., Pavelka, T.: Automatic Speech Recognition Using Context-dependent Syllables. Proceedings of 9th International PhD Workshop on Systems and Control (YGV2008), Izola, Slovenia, 2008, ISBN 978-961-264-003-3.

[3] Pavelka, T.: Hybrid Methods of Automatic Speech Recognition. PhD Thesis, University of West Bohemia, Pilsen, 2008.

Číslo aktivity

05/08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj rozpoznávače JLASER

Zahájení aktivity

4.1.2008

Ukončení aktivity

20.12.2008

Popis aktivity

Automatický rozpoznávač řeči JLASER se v současné době používá ve verzi 1.2, do níž byl nově přidán N-best dekodér, který umožní generování tzv. N-best listů obsahujících N nejpravděpodobnějších rozpoznaných sekvencí slov. Rovněž byl přidán modul automatického vyhodnocování úspěšnosti rozpoznávání, který kromě poměrně přesného měření úspěšnosti rovněž umožňuje výpočet intervalů spolehlivosti.

Skutečné Indikátory dosažení - výsledky aktivity

Zdrojové kódy programu (JLASER v.1.2) jsou dostupné na webových stránkách
http://liks.fav.zcu.cz/mediawiki/index.php/JLASER.
Program byl uveřejněn pod licencí GPL v.2.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Nový N-Best dekodér byl poměrně detailně popsán v článku:

[1] Pavelka, T., Brychcín, T.: N-Best Decoder for the JLASER Automatic Speech Recognizer. Proceedings of 9th International PhD Workshop on Systems and Control (YGV2008), Izola, Slovenia, 2008, ISBN: 978-961-264-003-3.

Číslo aktivity

06/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Vytvoření mapy slovních kategorií pro pořízené korpusy a návrh vhodného „neuronového“ systému pro kategorizaci dokumentů

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

V rámci aktivity byly dokončeny testy, natrénování a detailní rozbor sémantického obsahu mapy slovních kategorií založené na Kohonenově samoorganizující mapě. Zaměřili jsme se na syntaktickou a především sémantickou podobu jednotlivých slovních kategorií a dále pak na výslednou podobu mapy dokumentů (druhá vrstva architektury WEBSOM) vztaženou k podobě mapy slovních kategorií. Manuálně byla ohodnocena mapa slovních kategorií pro množinu 100 dokumentů z databáze ČTK (celkem 437 slovních kategorií). Dále byl implementován „batch“ algoritmus trénování Kohonenovy mapy a tento algoritmus byl začleněn do stávajícího systému. Algoritmus umožnil značné zrychlení trénování mapy slovních kategorií i následné mapy pro kategorizaci dokumentů. Další činnost v rámci této aktivity byla zaměřena na návrh a implementaci mapy dokumentů, a to neuronovou sítí umožňující kategorizaci dokumentů na základě informace o slovních kategoriích. Jako základní architektura byla pro mapu dokumentů zvolena opět Kohonenova samoorganizující mapa (podobně jako v případě systému WEBSOM). Vstupem mapy dokumentů byl zvolen modifikovaný vektor získaný z výstupu mapy slovních kategorií. Mapa dokumentů byla nejprve trénována množinou 100 dokumentů, v současné době probíhají testy s trénovací množinou o velikosti řádově několika tisíc dokumentů. Výsledky ukázaly, že výstup získaný z mapy dokumentů není příliš vhodný pro kategorizaci, neboť se v mapě obtížně lokalizují oblasti odpovídající jednotlivým třídám dokumentů. Z toho důvodu byla Kohonenova mapa nahrazena neuronovou sítí ART-2, která je opět trénována bez učitele, jejím výstupem je v porovnání s Kohonenovou mapou pouze jednoduchý vektor, ve kterém se mnohem snadněji stanovují a lokalizují jednotlivé kategorie. Další výhodou sítě je mnohem rychlejší trénování než u Kohonenovy mapy a možnost relativně snadného dotrénování sítě v případě přidávání nových dokumentů do korpusu. Nevýhodou sítě je obtížnější nastavení parametrů, které ovlivňují vlastní kategorizaci. Síť ART-2 byla implementována, začleněna do stávajícího systému a natrénována množinou o velikosti několika stovek dokumentů. Výsledky kategorizace a porovnání s mapou dokumentů na bázi Kohonenovy mapy byly publikovány v [2].

Skutečné Indikátory dosažení - výsledky aktivity

Manuální ohodnocení 437 slovních kategorií, výsledné souhrny zahrnující syntaktické a sémantické vazby v rámci jednotlivých slovních kategorií, vyhodnocení výsledné podoby mapy dokumentů vzhledem k výsledné mapě slovních kategorií, implementace "batch" algoritmu trénování a jeho začlenění do systému, implementace a testování neuronových sítí vhodných pro kategorizaci dokumentů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Byla vyhodnocena statistika úspěšnosti kategorizace dokumentů pro testované neuronové sítě a výsledky byly publikovány v následujících příspěvcích:

[1] Mouček, R., Mautner, P.: Sémantika přirozeného jazyka a reálného světa – počítačové zpracování. Sborník 4. ročníku mezinárodní konference „Informatika v škole a v praxi“, Ružomberok, Slovensko, 2008, ISBN

978-80-8084-362-5.
[2] Mautner, P., Mouček, R.: Zpracování a kategorizace česky psaných textových dokumentů neuronovou sítí, Sborník 4. ročníku mezinárodní konference „Informatika v škole a v praxi“, Ružomberok, Slovensko, 2008, ISBN 978-80-8084-362-5.
[3] Mouček, R., Mautner, P.: Categorization of Czech written documents using WEBSOM methods. In: Proceedings of 9th International PhD workshop on Systems and Control, Slovenia, 2008, ISBN 978-961-264-003-3.

Číslo aktivity
07/08
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Sumarizace textů založená na tenzorové LSA
Zahájení aktivity
2.1.2008
Ukončení aktivity
30.9.2008
Popis aktivity
V rámci této aktivity byl vytvořen sumarizátor založený na tensorové latentní sémantické analýze (LSA). Cílem bylo rozšířit maticovou metodu sumarizace shluku dokumentů založenou na LSA. Místo vstupní matice termů proti větám je zde použit tenzor termy/věty/dokumenty. Předpokládalo se, že dojde ke zpřesnění tvorby témat díky faktu, že podobné termy budou provázány jak společným výskytem ve větě, tak společným výskytem v dokumentu. Došlo však ke značnému nárůstu výpočetních a především paměťových nároků. K potlačení těchto nároků byla nasazena metoda náhodného indexování vektorů tenzoru, která je schopna výrazně snížit jeden rozměr tenzoru. Avšak jejím vedlejším efektem bylo vnesení prvku náhody do sumarizačního procesu, které způsobilo, že při každém běhu sumarizátoru se stejným nastavením se mohly vybrat do souhrnu jiné věty. Tento problém značně ztěžoval zjištění optimálního nastavení sumarizátoru. Existovala zde možnost řešit negativní vliv náhody tím, že sumarizátor bude spuštěn vícekrát a metodou hlasování budou do souhrnu vybrány ty věty, které se budou vyskytovat v největším počtu běhů. To by ovšem opět prodloužilo potřebný čas. Experimenty navíc ukázaly, že nedošlo ke zlepšení kvality souhrnů ve srovnání s maticovou metodou. Proto byla tenzorová metoda opuštěna a maticová metoda dále rozpracována. Maticový LSA model byl rozšířen o metodu Iterative Residual Rescaling, která umožňuje zmírnit vliv zbytkových vektorů dominantních témat při tvorbě latentního prostoru. Navíc byla modifikována metoda výběru vět do souhrnu.
Skutečné Indikátory dosažení - výsledky aktivity
V rámci řešení byl vytvořen experimentální systém, který umožňuje sumarizovat shluk dokumentů týkajících se určité události/tématu. Lze použít jak tenzorovou metodu, tak rozpracovanou maticovou metodu. Idea tenzorové metody byla zmíněna v [1]. Vstupem sumarizátoru je xml soubor navržený a popsaný v předchozích aktivitách tohoto projektu.

Přikládáme porovnání tensorové a maticové verze LSA sumarizátoru (porovnání na kolekci anglických textů DUC 2005, ROUGE metoda hodnocení):

ROUGE-1 ROUGE-2
Maticová LSA sumarizace 0,32759 0,06250
Tenzorová LSA sumarizace 0,33106 0,05899

Z pohledu skóre ROUGE-1 je lepší tenzorová metoda, z pohledu ROUGE-2 je to opačně. Rozdíly však nejsou statisticky významné. Když jsme vzali v úvahu problémy při použití tenzorové metody rozebrané v popisu aktivity, rozhodli jsme se pokračovat v další fázi s maticovou metodou.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Byl zpracován a ověřen experimentální sumarizační systém založený na tenzorové latentní sémantické analýze. Systém byl popsán v publikaci:
[1] Ježek, K., Steinberger, J.: Automatic Text Summarization (The state of the art 2007 and new challenges). In: Proceedings of Znalosti 2008, Bratislava, Slovakia, 2008, pp. 1–12, ISBN 978-80-227-2827-0.

Číslo aktivity
08/08
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Řazení vět v souhrnu
Zahájení aktivity
1.2.2008
Ukončení aktivity
30.9.2008
Popis aktivity
V případě sumarizace jednoho dokumentu jsou věty v souhrnu řazeny dle jejich výskytu v původním textu. Při přechodu k sumarizaci shluku dokumentů týkajících se stejné události/tématu se objevil nový problém. Věty přicházející do souhrnu z různých dokumentů musejí být seřazeny tak, aby výsledný dokument vypadal jednotně a aby věty na sebe co nejlépe navazovaly. Cílem bylo tedy navržení a otestování metody, která seřadí věty v souhrnu. Myšlenkou bylo, aby věty, které se vyskytnou vedle sebe ve výsledném souhrnu, obsahovaly co nejvíce výskytů společných entit. Další vlastnost, kterou jsme mohli použít, byl datum vzniku dokumentů, ve kterých se jednotlivé věty nacházely. Navržený algoritmus byl implementován jako modul sumarizátoru. Dále jsme provedli pilotní experiment s kolekcí DUC2007. Sumarizátorem byly nejprve vytvořeny souhrny. Tyto souhrny byly dále anotovány – věty byly seřazeny ručně a také seřazeny navrženou metodou. Nakonec se zjišťovala korelace mezi ručním seřazením a automatickým.
Skutečné Indikátory dosažení - výsledky aktivity
V rámci aktivity byl vytvořen modul sumarizátoru řadící věty v souhrnu. Lze spustit dávkově (vstupem je xml soubor, který již obsahuje souhrny) nebo jej lze začlenit do sumarizátoru (viz aktivita č. 07/08).

Výsledky korelace mezi pořadím vět souhrnu anotovaným ručně a pořadím vzniklým automatickou metodou s následujícími nastaveními:
1. Baseline 1 – věty řazeny dle jejich vzájemné kosinové podobnosti (začíná se nejpodobnější dvojicí vět, pokračuje se nejpodobnější větou k druhé větě atd.).
2. Baseline 2 – použit pouze datum dokumentu, věty ze stejného dokumentu jsou řazeny dle výskytu v dokumentu.
3. Řazení dle výskytu entit.
4. Kombinované řazení dle výskytu entit a data dokumentu.

Délka souhrnů byla 100 slov a 250 slov. Rozdíly korelací jednotlivých metod mezi 2, 3 a 4 nebyly statisticky významné, systém 1 byl statisticky významně horší než kterýkoliv systém z množiny 2, 3 a 4.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Byl zpracován softwarový modul sumarizačního systému – řazení vět v souhrnu. Modul bude po řádném otestování dán k dispozici. Je zkušebně dostupný v rámci systému SWEeT na URL:
http://tmrg.kiv.zcu.cz:8080/sweet/

Číslo aktivity
09/08
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Online vyhledávací a sumarizační systém
Zahájení aktivity
1.2.2008
Ukončení aktivity
22.12.2008
Popis aktivity
V rámci této aktivity byl vytvořen systém, který pracuje následovně: Uživatel vloží dotaz, který by měl být dostatečně bohatý, aby vymezil dané téma. Systém následně vyhledá nejrelevantnější dokumenty k vloženému dotazu. Pak je spuštěn řetěz sumarizačních modulů, jehož cílem je vytvořit souhrn vyhledaných dokumentů. Je zde použit vyvíjený sumarizátor. Výsledný souhrn je potom vrácen uživateli jako odpověď. V další fázi projektu bude systém rozšiřován o další moduly, které by měly zlepšit kvalitu vytvořených souhrnů.
Skutečné Indikátory dosažení - výsledky aktivity
Byl vytvořen on-line vyhledávací a sumarizační systém. Jeho architektura a popis jednotlivých částí lze najít v [2]. Systém je dostupný na webu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
Vyhledávací a sumarizační systém SWEeT je testován a zkušebně dostupný na URL:
http://tmrg.kiv.zcu.cz:8080/sweet/

[1] Steinberger, J., Ježek, K., Sloup, M.: Web Topic Summarization. In: Proceedings of the 12th International Conference on Electronic Publishing, pp 322-334, Toronto, Canada 2008, ISBN 978-0-7727-6315-0.

Číslo aktivity
10/08
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Experimenty v systému pro vyhledávání
Zahájení aktivity
2.1.2008
Ukončení aktivity
31.12.2008
Popis aktivity
V rámci projektu navrhujeme prototypové řešení multilingválního vyhledávání obohacené o automatickou sumarizaci vyhledaných textů. Jádrem vyhledávání je thesaurus EuroWordNet a sumarizátor je založen na latentní sémantické analýze. Současné řešení obsahuje především možnosti zpracování anglického a českého jazyka. V rámci aktivit předchozích let jsme provedli implementaci systému pro rozšiřování dotazů a disambiguaci (aktivity 2007-17 a 2007-18). V roce 2008 jsme prováděli testy na celém prototypovém systému. Pro testy byl použitý vícejazyčný korpus textu Evropské unie – JRC-EU a korpus Fairy-tale se zjednodušenou slovní zásobou. Sledována byla přesnost a úplnost zpracování v obou případech navrženými metodami. Zahrnutím vícejazyčného zpracování s využitím tezauru EWN bylo umožněno křížové zpracování. Ve druhé části jsme se zaměřili na modul

pro odhalování plagiovaných částí textu, což je detailně popsáno a publikováno v citovaném článku [1].

Skutečné Indikátory dosažení - výsledky aktivity

Navržený systém byl testován z hlediska relevance a úplnosti vyhledávaných dokumentů při aplikaci jednotlivých modulů. Byla zajištěna a testována především jejich součinnost a vliv na obdržené výsledky. Jako rozsíření nad rámec plánovaného rozsahu byl navržen modul pro detekci plagiovaných příspěvků. Detekována je redundantní informace, což následně umožňuje snazší orientaci ve vyhledaných dokumentech.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

U systému jsme prováděli testy určující relevanci získaných výsledků jak pro české a anglické prostředí, tak i při křížovém zpracování. Provedeno bylo také srovnání výsledků s přístupem aplikovaným ve vyhledávači Google. Systém jsme dále testovali na korpusech textů Evropského parlamentu s důrazem na detekci redundantních částí textu. Výstupem aktivity je článek publikovaný na konferenci AIMSA 2008.

[1] Ceska, Z., Toman, M., Jezek, K.: Multilingual Plagiarism Detection. Proc. of the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2008), Varna, Bulgaria, September 2008, LNCS/LNAI 5253, pp. 83-92, Springer-Verlag Berlin, Heidelberg, September 2008. ISSN 0302-9743, ISBN 978-3-540-85775-4.

Číslo aktivity

11/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Extrakce informace z webových sídel

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.9.2008

Popis aktivity

Prostředí webu se postupně vyvinulo v obecný zdroj informací uchovávaných převážně v částečně strukturovaném formátu HTML. Stávající Web obsahuje data, která jsou určena pro prohlížení uživatelem, jenž zobrazenému textu přiřadí správnou sémantickou informaci. Jednotlivé zdroje vyjadřují informace v rozdílných formátech a různým způsobem, což pro člověka nepředstavuje větší problém, ale komplikuje jejich porozumění počítačem. Taková situace skýtá velké množství netriviálních úloh, které ve výsledku mohou vést k transformaci stávajících zdrojů do tzv. sémantického webu. Jednou větví výzkumu v této oblasti je extrakce informací z dat (IE Information Extraction). Cílem extrakce dat je v našem případe získat z webových stránek čistý text a přidružená metadata, např: čas publikování příspěvku, autora, kategorii, název, perex. Vlastní text článku je však v praxi velmi často nesouvislý a je přerušen multimediálními daty, případně reklamou, což představuje komplikace.

Skutečné Indikátory dosažení - výsledky aktivity

Navržený systém umožňuje extrakci vybraných dat z webových stránek. První námi navržená metoda je založena na statistické analýze struktury webové stránky a druhá metoda využívá dotazy XQuery pro extrakci informace z částečně strukturovaných dokumentů. V testech srovnáváme přesnost a úplnost automatické extrakce pomocí obou metod a ručně vytvářeného referenčního extraktu. Metoda XQT produkuje přesné výsledky s možností jemného strukturování výsledných dat. Je vhodná pro tvorbu textových korpusů a extrakci obecných dat s důrazem na přesnost. Metoda NIT poskytuje výsledky s přesností a úplností pohybující se kolem 80 %. Použití metody je jednoduché, protože kvalita extrakce závisí na jediném parametru. Metoda je určená výhradně pro tvorbu textových korpusů z webových zdrojů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

V rámci řešení jsme navrhli dvě alternativní metody pro extrakci informace z webu. Výstupem aktivity je článek popisující algoritmus extrakce spolu s výsledky. Článek byl publikován ve sborníku konference Znalosti 2008.

[1] Toman, M.: Srovnání přístupů extrakce užitečné informace
z webu, Sborník konference Znalosti 2008, Bratislava,
Slovakia, 2008, ISBN 978-80-227-2827-0.

Číslo aktivity

12/08

Ke kterému dílčímu cíli se aktivita vztahuje

1 - Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřová...

Název (cíl)aktivity

Příprava a vytvoření datové kolekce pro testování dotazů nad sémantickým webem

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

Výsledkem této aktivity bylo vytvoření dvou datových kolekcí. První datová kolekce byla vytvořena řešitelem ručně a obsahuje aktuální informace o studijních programech a oborech technických vysokých škol a univerzit v ČR. Do kolekce jsou zahrnuty pouze ty studijní programy a obory, které se zabývají problematikou computer science nebo softwarovým inženýrstvím. Jako podkladové materiály pro tvorbu kolekce posloužily dokumenty z internetových stránek MŠMT (číselník studijních programů STUDPROG), Českého statistického úřadu (klasifikace kmenových oborů vzdělávání KKOV), portály www.scio.cz a www.vysokeskoly.cz, a internetové stránky jednotlivých vysokých škol a univerzit. Řešitel tuto kolekci začal doplňovat též i předměty, které se v daných studijních oborech vyučují. Tato činnost byla časově velmi náročná a nakonec kolekce obsahuje jen předměty jedné vybrané fakulty, konkrétně FAV ZČU v Plzni. Druhá datová kolekce byla vytvořena elektronickou cestou. Student M. Dostal vytvořil webového robota, který stáhnul a do databáze uložil během dvou měsíců přibližně 9 000 článků týkajících se katastrof. K tomuto účelu byl použit web www.katastrofy.com. Obsah databáze byl následně exportován do formátu XML.

Skutečné Indikátory dosažení - výsledky aktivity

Byly vytvořeny dvě datové kolekce ve formátu XML. První kolekce obsahuje data získaná z 24 fakult z 21 různých technických vysokých škol a univerzit v ČR. Celkem datová kolekce čítá 31 studijních programů (bakalářské, magisterské, navazující magisterské a doktorské) a 93 studijních oborů (bakalářské, magisterské a doktorské – prezenční i kombinované studium). V kolekci je podchycena skladba každého studijního programu, tj. jaké obory v rámci tohoto studijního programu lze studovat. Jedna fakulta (konkrétně FAV ZČU v Plzni) obsahuje navíc 165 předmětů s anotacemi. Tyto předměty jsou použity jako náplň studijních oborů na této fakultě, u každého předmětu je uveden status zápisu na daný obor. Druhá kolekce obsahuje přibližně 9000 článků na téma katastrofy.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vytvořené datové kolekce jsou po dohodě ke stažení ze stránek textmining.zcu.cz (kontakt zima@kiv.zcu.cz).

Číslo aktivity

13/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Transformace ontologie na logický pravidlově orientovaný program a zpět

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

V tomto roce byla navržena a realizována jednoduchá ontologie nad datovou kolekcí katastrof, která byla vytvořena v rámci aktivity 2008-12. Pro zápis ontologie byl definován jazyk založený na principu trojic (objekt, predikát, předmět, např. ‘katastrofa‘, ‘subtype‘, ‘povodeň‘) a využívá podobných vlastností jako jazyky RDF, resp. RDFS. Výsledná ontologie byla aplikována v úloze sémantického vyhledávání příslušných článků na základě položeného dotazu (např. kdy a kde vznikl požár v roce 2006). Aplikace využívá data uložená v databázi, která mohou být webovým robotem aktualizována. Celá aplikace je napsána v jazyce PHP, data jsou uložena v databázovém systému MySQL. Hlavním cílem této aktivity je transformace navržené ontologie do logického programu, založeného na pravidlově orientovaném jazyce, a též zpětná transformace. Rok 2008 byl prvním rokem řešení této aktivity, v jejím řešení se bude pokračovat i v následujícím roce v rámci navržených aktivit „Transformace zvolené ontologie do logického programu v jazyce Datalog“ a „Zobecnění transformace ontologie do logického programu v jazyce Datalog“.

Skutečné Indikátory dosažení - výsledky aktivity

Byla vytvořena ontologie nad datovou kolekcí katastrof, webová aplikace založená na sémantickém vyhledávání a aplikace byla otestována na reprezentativní množině dotazů.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Webová aplikace sémantického vyhledávání je dostupná na http://tmrg.kiv.zcu.cz/semweb.

Číslo aktivity

14/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Editor pro anotaci korpusu dialogovými akty

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

V rámci této aktivity, navazující na 2006-14, byl vytvořen nástroj, který umožňuje anotaci korpusu dialogovými akty. Hlavní obrazovka aplikace zobrazí text k anotaci (dialog ve formátu txt nebo xml). Pomocí myši je možno označit oblast k anotaci, ke které je posléze doplněna značka dialogového aktu. Důraz je zde kladen na rychlost anotace: značka anotovaného dialogového aktu se přidává stiskem klávesové zkratky.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikaci, pomocí níž je možné anotovat korpusy dialogovými akty, je možné stáhnout z adresy http://home.zcu.cz/~pkral/DAlabel.zip .

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Aplikace byla ověřena při manuální anotaci korpusu. Ručně bylo anotováno celkem 1518 dialogových aktů.

Číslo aktivity

15/08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Automatické rozpoznávání dialogových aktů

Zahájení aktivity

10.4.2008

Ukončení aktivity

15.12.2008

Popis aktivity

Tato aktivita se zabývala automatickým rozpoznáváním dialogových aktů a navazovala na aktivity 2006-13 a 2007-33. V rámci těchto předchozích aktivit byly využity k rozpoznávání dialogových aktů dva hlavní zdroje informací: lexikální informace a prozódie. Některé studie však ukazují, že je výhodné tyto informace doplnit o tzv. dialogovou historii (časovou sekvenci po sobě jdoucích dialogových aktů). Proto jsme se nyní zaměřili na prostudování dostupných metod, které tuto informaci využívají. Studie potvrzují, že dialogová historie může zvýšit přesnost rozpoznávání dialogových aktů, v našem případě se ale ukázalo, že toto zvýšení nebude významné. Naše nejlepší metoda "Best Position", která využívá informaci o struktuře věty, již totiž dosahuje 96 % přesnosti (v kombinaci s prozodií 97 %). Rozhodli jsme se proto dialogovou historii do našeho systému nezahrnout. Dále jsme se zaměřili na návrh dalších nových metod rozpoznávání dialogových aktů a vytvořili dvě nové metody: „Interpolated Multiscale Position" a „Frequency Bin Interpolation". Obě metody jsou rozšířením naší předchozí metody „Multiscale position", avšak místo Back-off přístupu používají lineární interpolaci.

Skutečné Indikátory dosažení - výsledky aktivity

Na určení aktuálního dialogového aktu na základě dialogové historie se používá celá řada statistických i jiných metod. Zjistili jsme, že nejlépe fungují statistické metody, a to: n-gramy, skryté Markovovy modely a dynamické Bayesovy sítě. Experimenty ukázaly, že všechny tyto metody fungují uspokojivě a že je možné použít jakoukoli z nich. Každý z autorů používá většinou jiný korpus a jinou množinu dialogových aktů, takže je velmi složité rozhodnout, která ze statistických metod je nejlepší.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Popisem a srovnáním metod využívajících dialogovou historii se zabývá článek s názvem „Dialogue Act Recognition Approaches", který byl v červenci podán k recenzi do impaktovaného časopisu Computing and Informatics:

Kral, P. and Cerisara, C.: Dialogue Act Recognition Approaches. In: Computing and Informatics, Slovenská akademie věd, 2008/09, (podáno 07/2008, zatím není znám výsledek recenze).

Nové metody automatického rozpoznávání dialogových aktů byly publikovány na prestižní mezinárodní IEEE konferenci Machine Learning for Signal Processing 2008:

Kral, P., Pavelka, T., Cerisara, C.: Evaluation of Dialogue Act Recognition Approaches. In: MLSP""08, Cancun, Mexico, October 2008, pp. 492 - 497, ISSN: 1551-2541, ISBN: 978-1-4244-2375-0.

Číslo aktivity

16/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Tvorba českého korpusu plagiátů textových dokumentů

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.6.2008

Popis aktivity

V rámci této aktivity byl manuálně vytvořen korpus obsahující dokumenty, které jsou s různým procentuálním podílem okopírovány z jiných textových dokumentů. Výsledný korpus 1500 textových dokumentů je uložen v plain-text formátu, každý soubor je označkován. Soubory plagiátů (celkový počet 550) mají název ve tvaru

„query_number source_1,source_2,…“, kde query_number označuje unikátní číslo plagiátu začínající písmenem „q“ (s následujícím pětimístným číselným kódem). Identifikátor source_x je šestimístné unikátní číslo zdrojového dokumentu, ze kterého byl plagiát vytvořen. Každý plagiát může být vytvořen z jednoho i více zdrojů. Zdrojové dokumenty o politice (v počtu 300) a další články podobných témat (v počtu 650) jsou označeny názvem ve tvaru „unique_id nazev“, kde unique_id je šestimístné unikátní číslo odpovídající identifikátoru článku v ČTK. Struktura textu v dokumentech je následující: Každý odstavec textu je na samostatné řádce, mezi jednotlivými odstavci je jedna prázdná řádka. Věty v rámci jednoho odstavce následují za sebou na jedné řádce a jsou korektně ukončeny tečkou. Textové dokumenty plagiátů byly vytvořeny manuálně studenty dle stanovaných kritérií. Výsledný text je získán použitím jednoho či více zdrojů a vložením vlastních myšlenek. Při vytváření plagiátu byly použity tyto postupy: a) Výměna odstavců, vět, několik po sobě následujících slov (častá situace, cca 40%) b) Smazání celé věty, její části nebo jednoho slova (obvyklá situace, cca 25%) c) Výměna jednoho slova nebo náhrada synonymem (obvyklá situace, cca 15%) d) Vložení vlastních slov pro napojení významu věty (méně obvyklá situace, cca 10%) e) Prohození pořadí odstavců, vět, slov (méně obvyklá situace, cca 10%)

Skutečné Indikátory dosažení - výsledky aktivity

Český korpus plagiátů obsahující 1500 textových dokumentů o politice. Na tomto korpusu bylo provedeno testování vytvářených algoritmů - viz aktivita č. 18.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

U vytvořeného korpusu jsme provedli vizuální kontrolu. Výsledný korpus slouží jako podklad pro aktivitu č. 18. Korpus se sestává z 1500 textových dokumentů o politice, kde 550 článků je plagiovaných (manuálně vytvořených studenty) a 300 článků jsou zdrojové podklady, z kterých vychází plagiáty. Dalších 650 článků je náhodně vybráno z oblasti polity. Tyto články se týkají podobných témat jako plagiáty, nicméně přímo nesouvisí se zdrojovými podklady. Korpus s jeho popisem je dispozici na webových stránkách http://textmining.zcu.cz/public/NPV/2008/aktivita16/KorpusPlagiaty.zip.

Číslo aktivity

17/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Návrh a vývoj úložiště textových dokumentů

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Pro uchování korpusu plagiátů s možností centralizovaného přístupu jsme navrhli datové úložiště. Spolu s úložištěm jsme vytvořili aplikaci pro import textů a dokumentů ve specializovaném formátu, jenž se používá např. v ČTK korpusu. Tato aplikace usnadňuje hromadné plnění databáze experimentálními daty a testování metod odhalující plagiáty. Import je zaměřen na vložení korpusu z aktivity č. 16 a dalších textových dokumentů pro výzkumné potřeby. Nástroj pro import spolu se skripty pro vytvoření datového úložiště jsou k dispozici na webových stránkách http://textmining.zcu.cz/public/NPV/2008/aktivita17/UlozisteTxtDok.zip. Nástroj pro import je zpracován formou konzolové aplikace, která byla vyvinuta v prostředí .NET Framework 2.0. Datové úložiště využívá Microsoft SQL Server 2005.

Skutečné Indikátory dosažení - výsledky aktivity

Datové úložiště schopné uchovat textové dokumenty, umožňující jejich třídění a rychlé procházení. Funkční aplikace schopná importovat textové dokumenty do datového úložiště.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

U aplikace pro import dat jsme provedli vizuální kontrolu a ruční otestování. Spolu s tím jsme testovali datové

úložiště, kam byla importována předem specifikovaná data.

Číslo aktivity

18/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Implementace a experimentální ověření metody odhalující plagiáty textových dokumentů s využitím latentní sémantické analýzy

Zahájení aktivity

2.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Z důvodu dosud nízkého zájmu o ověřování původu studentských prací jsme implementovali metodu s optimalizací pro české prostředí. Do výběru je též zahrnuta podpora pro anglický jazyk. Aktivita volně navazuje na předchozí aktivitu č. 36 z roku 2007, která se zabývala návrhem experimentální metody pro identifikaci plagiátů textových dokumentů s využitím latentní sémantické analýzy. V rámci této aktivity jsme implementovali a ověřili metodu pro odhalování plagiátů v psaném textu, která je založena na latentní sémantické analýze. Tato metoda využívá singulární dekompozice vztahů frází zastoupených ve zkoumaných dokumentech. Pro redukci množství frází jsme navrhli důmyslný filtr, který uvažuje významnost frází a odstraňuje nepotřebné fráze. Díky tomu se nám podařilo podstatně eliminovat objem dat vstupující do latentní sémantické analýzy. Výsledky této metody, včetně porovnání s existujícími přístupy, byly publikovány na konferencích GoTAL 2008 a ITAT 2008. Jako zdrojová data pro experimenty jsme použili korpus z aktivity č. 16. Kromě toho jsme se zabývali možnostmi vícejazyčného prostředí, konkrétně češtinou a angličtinou. Výsledky vícejazyčného zpracování byly publikovány na konferenci AIMSA 2008.

Skutečné Indikátory dosažení - výsledky aktivity

Ověřená metoda pro detekci plagiátů využívající latentní sémantickou analýzu. Výsledky experimentů byly publikovány na řadě významných konferencí.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledkem aktivity jsou publikace:

• Ceska, Z.: „Plagiarism Detection based on Singular Value Decomposition“. Advances in Natural Language Processing, LNCS/LNAI 5221, pp. 108-119, Springer Verlag Berlin Heidelberg, the 6th International Conference on Natural Language Processing (GoTAL 2008), Gothenburg, Sweden, August 2008. ISSN 0302-9743. ISBN 978-3-540-85286-5.

• Ceska, Z., Toman, M., Jezek, K.: „Multilingual Plagiarism Detection“. Artificial Intelligence: Methodology, Systems, and Applications, LNCS/LNAI 5253, pp. 83-92, Springer-Verlag Berlin Heidelberg, the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2008), Varna, Bulgaria, September 2008. ISSN 0302-9743. ISBN 978-3-540-85775-4.

• Ceska, Z.: „Využití moderních přístupů pro detekci plagiátů“. Proceedings of the ITAT 2008, Information Technologies – Applications and Theory, Hrebienok, Slovakia, pp. 23-26, September 2008. ISBN 978-80-969184-8-5.

• Ceska, Z.: „Free-Text Plagiarism Detection Based on Latent Semantic Analysis“. Technical Report No. DCSE/TR-2008-01, Pilsen, Czech Republic, April 2008.

Číslo aktivity

19/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Optimalizace a finální implementace metody Teraman

Zahájení aktivity

2.1.2008

Ukončení aktivity

30.9.2008

Popis aktivity

Tato aktivita volně navazovala na aktivitu č. 22 z roku 2007, v jejímž rámci jsme navrhli metodu (nástroj) Teraman pro extrakci N-gramů z rozsáhlých textových dat. U navržené metody jsme provedli optimalizaci v části dávkového zpracování souborů a implementaci knihovních funkcí. Finální implementace nástroje Teraman spolu s optimalizovaným algoritmem je k dispozici na webových stránkách http://textmining.zcu.cz/public/NPV/2008/aktivita19/Teraman.zip. Aplikace je formou DLL knihovny vyvinuté v prostředí .NET Framework 2.0, kterou lze využít jako COM+ knihovnu. Součástí je rovněž konzolová aplikace pro spouštění z příkazové řádky a stručný návod pro použití. Výsledky dosažené realizací této aktivity jsou taktéž výše uvedené články.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace schopná akceptovat definované vstupy a poskytnout požadované výstupy. Možnost různých nastavení nutných pro experimenty a provoz na různých platformách.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vizuální kontrola a ruční otestování aplikace, použití předem daných dat a následná kontrola očekávaného výsledku.

Kromě toho je výsledkem aktivity publikace:

• Ceska, Z., Hanak, I., Tesar, R.: „Extrakce N-gramů z rozsáhlých textů". In: Proceedings of the 7th Annual Conference ZNALOSTI 2008, Bratislava, Slovakia, pp. 54-65, February 2008. ISBN 978-80-227-2827-0.

Číslo aktivity

20/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

SPOT – nový webový projekt on-line slovníku překladů odborných termínů

Zahájení aktivity

2.1.2008

Ukončení aktivity

Popis aktivity

Implementace slovníku dospěla do stadia realizace základní funkcionality, rutinním provozem byla ověřena webová aplikace slovníku. V průběhu roku byla doplněna o funkčnost hromadného importu slovníkového korpusu (zatím v ověřovacím provozu) a slovník byl naplněn korpusem, jehož autorem je spoluřešitel aktivity J.Hynek. V této podobě je aplikace dostupná jako softwarový prototyp pod licencí Creative Commons. V rámci ekosystému projektu je poměrně úspěšný web blogspot.zcu.cz s postupně narůstající návštěvností, která představuje cca 16500 návštěvníků za měsíce 10+11/2008. Aktuální vývoj je směřován na odstranění nedostatků aplikace nalezených provozem a zlepšení jejího ovládání. Implementace jsou prováděny v kontextu bakalářské práce (T.Peterka, vedoucí P.Brada, obhájena v červnu 2008) a semestrálního projektu. Doprovodný portál www.blogspot.cz je řešen v kontextu semestrálního projektu PRJ5 (M. Homolka, vedoucí J. Hynek) a navazující bakalářské práce. Dokončením implementace vznikne zatím jediná tuzemská platforma profesionálních překladatelů s možností

efektivního ustalování nové či řídce užívané odborné terminologie. Výstupy této části portálu budou po redakčních úpravách sloužit k pravidelné aktualizaci SPOTu. SPOT bude rovněž přímo přístupný z tohoto portálu formou portletu. Další funkcionalita zahrnuje odborné profily uživatelů, týmové projekty, odborný blog, diskusní skupiny, tematické články, sdílení specializovaných glosářů a další. Cílem pro následující období je rozšíření uživatelské základny včetně integrace slovníku do jiných aplikací, sběr zkušeností s takovýmto typem slovníku, a dále plnohodnotná podpora pro překladatelské projekty. Detaily jsou dostupné na http://wiki.kiv.zcu.cz/SlovnikTerminologie/HomePage.

Skutečné Indikátory dosažení - výsledky aktivity

Funkční aplikace dostupná na http://spot.zcu.cz/ naplněná korpusem. Rutinní provoz s pilotní množinou uživatelů.
Doprovodný blog k výše uvedenému na adrese http://www.blogspot.cz (cca 70 tisíc přístupů za 1-11/2008, dle údajů Active 24 webhosting).
Obhájená bakalářská práce Tomáš Peterka: Použití rámců Java EE pro webové aplikace.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Aplikace SPOT registrovaná jako autorizovaný software (typ výsledku S) v evidenci RIV.
Funkční aplikace dostupná na http://spot.zcu.cz/, naplněná korpusem.

Číslo aktivity

21/08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh a implementace TTS Systému pro syntézu českého jazyka z textu

Zahájení aktivity

1.6.2008

Ukončení aktivity

15.12.2008

Popis aktivity

Pro komunikaci pomocí přirozeného jazyka je potřeba nejen rozpoznat a interpretovat řečový signál uživatele, ale též předat odpověď systému uživateli v podobě akustického signálu (ne v podobě textu). Proto vznikla tato aktivita, která se zabývá text-to-speech (TTS) syntézou, tj. vytvořením řečového signálu z textu pomocí počítače. Na systém jsou kladeny dva následující požadavky: (1) vzniklý signál musí být v češtině (2) syntetizovaná řeč musí být přirozená.

Skutečné Indikátory dosažení - výsledky aktivity

V jazyce Java byl vytvořen jSynt TTS systém, který pro syntézu řeči používá systém MBROLA. Aktuální verze systému podporuje syntézu dvou jazyků: češtinu a angličtinu. Syntetizovaná řeč je srozumitelná, není ale příliš přirozená. Větší přirozenosti syntetizované řeči bude v budoucnosti možné docílit úpravou či nastavením dalších parametrů systému.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Vytvořený systém byl publikován na mezinárodní konferenci v následující publikaci:
Kral, P., Ekstein, K.: jSynt: A Czech Text-to-Speech System written in JAVA. In: 9th International PhD Workshop on Systems and Control, Izola, Slovenia, October 2008, ISBN: 978-961-264-003-3.

Číslo aktivity

22/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Metody automatického rozpoznávání výrazu obličeje

Zahájení aktivity

15.1.2008

Ukončení aktivity

15.12.2008

Popis aktivity

Tato aktivita navazuje na aktivitu 2007-33 a zabývá se automatickým rozpoznáváním výrazu obličeje. Převážná část existujících metod automatického rozpoznávání přirozené řeči využívá k určení aktuálního dialogového aktu kombinaci lexikálních (většinou v podobě posloupnosti slov ve větě) a prozodických příznaků. Bohužel, prozodické příznaky jsou velmi často nedostatečné a je nutné doplnit je dalšími informacemi, např. výrazem obličeje. Velkou výhodou a v podstatě nejzajímavější vlastností vytvořeného systému v porovnání s existujícími systémy je plná automatizace, která je nezbytnou podmínkou pro komunikační rozhraní člověk – stroj. Jde o multiuživatelský systém pracující ve dvou módech: statickém - snímky a dynamickém – snímky a video.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem této aktivity je automatický uživatelsky nezávislý systém rozpoznávání výrazu obličeje – ARFE, který byl zadán do RIV – III. Kategorie, podkategorie S - autorizovaný software. Dále je přijat příspěvek na konferenci v roce 2009.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Byl zpracován a přijat příspěvěk na konferenci v roce 2009 - detaily příspěvku budou uvedeny v příštím roce.
Byl zprovozněn autorizovany software ARFE.

Číslo aktivity

23/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Získávání znalostí analýzou struktury Webu

Zahájení aktivity

1.3.2008

Ukončení aktivity

30.6.2008

Popis aktivity

Tato aktivita nebyla původně plánována a byla zařazena dodatečně proto, že časopisecká publikace vázající se k aktivitě 09/2007 byla vydána až v roce 2008 a k ní se vázající diskuse přinesly dokonalejší pohled na řešení problému. Byl navržen a zdůvodněn nový formální popis řešení a spolu s výsledky testů rozšířených o aktuální data pak byly publikovány na konferenci ELPUB 2008. Rovněž byly stanoveny další možnosti objektivizace metody, jejich implementace však závisí na získání vhodného výzkumníka.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem je funkční metoda, schopná objektivněji hodnotit významnost autorů než dovolují stávající bibliografické metody. Metoda byla uznána za hodnou časopiseckého publikování v periodiku specializovaném na hodnocení výzkumu.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Článek v impaktovaném časopisu:
Fiala D., Rousselot F., Jezek K.: PageRank for Bibliographic Network. Scientometrics, vol.76, no. 1, pp. 135-158, 2008-12-18, ISSN 0138-9130, Akademiai Kiado, Springer
Článek ve sborníku konference:
Fiala D., Jezek K., Steinberger J.: Exploration and Evaluation of Citation Network. In Proceedings of the 12th

International Conference on Electronic Publishing, ISBN 978-0-7727-6315-0, pp 351-362, Toronto, Canada 2008

Číslo aktivity

25/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Formalizace lexikální databáze VerbaLex obsahující valenční rámce českých sloves a jejich vazby na princetonský WordNet v.2.0 a v.3.0.

Zahájení aktivity

1.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Pro komplexní valenční rámce získané při budování databáze Verbalex byly hledány vhodné notační varianty použitelné jak v syntaktickém analyzátoru, tak i v aplikaci pro manipulaci se sémantickými rámci a reprezentacemi. V souvislosti s pracemi na světovém wordnetovém gridu byly analyzovány použitelné datové struktury a jejich reprezentace s ohledem na vazby s princetonským (a dalšími) Wordnetem. Byly též vytvořeny nástroje a postupy pro světový wordnetový grid.

Skutečné Indikátory dosažení - výsledky aktivity

Výsledkem jsou soubory rámců použitelných v experimentech s analyzátorem Synt a při experimentech s přístupy k webu v přirozeném jazyce. U vazeb na princetonský Wordnet byl vytvořen seznam translačních ekvivalentů napojených na ILI s využitím nástroje WordNet Assistent.

Naše platforma DEB a nástroj DEBVisDic byly zvoleny jako architektura pro novou aktivitu tvorby globální provázané sítě národních sémantických sítí typu WordNet, tzv. Global WordNet Grid (GWG).

DEB díky flexibilitě svého návrhu umožňuje v rámci GWG řešit

i otázky, které zatím bránily uvolnění přístupu k národním wordnetům (v současnosti jich existuje asi 50), jako jsou omezující licenční podmínky. V rámci GWG a DEB platformy je možné umístit data buď na centrální úložiště nebo na místní server poskytovatele s tím, že pro dotazy přistupující ke GWG se jeví tyto varianty jako rovnocenné. V současnosti obsahuje GWG 4 národní wordnety, další budou přidány ve vazbě na EU projekt KYOTO.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Horák, Aleš - Rambousek, Adam - Vossen, Piek: A Distributed Database System for Developing Ontological and Lexical Resources in Harmony. In 9th International Conference on Intelligent Text Processing and Computational Linguistics. Berlin : Springer, 2008. od s. 1-15, 15 s. ISBN 978-3-540-78134-9.

Horák, Aleš. Computer Processing of Czech Syntax and Semantics. 1st edition. Brno, Czech Republic : Librix.eu, 2008. 241 s. 1st edition. ISBN 978-80-7399-375-7.

Horák, Aleš - Rambousek, Adam - Maks, Isa - Segers, Roxane - Vossen, Piek - van der Vliet, Hennie. Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology. In The 18th International Congress of Linguists. Seoul, Republic of Korea : Korea University, 2008. od s. 190-191, 2 s.

Němčík, Václav - Hlaváčková, Dana - Horák, Aleš - Pala, Karel - Úradník, Michal. Processing Czech Verbal Synsets with Relations to English WordNet. In RASLAN 2008. 2. vyd. Brno : Masarykova Univerzita, 2008. od s. 49-55, 7 s. ISBN 978-80-210-4741-9.

Němčík, Václav - Pala, Karel - Hlaváčková, Dana. Semi-automatic Linking of New Czech Synsets Using Princeton WordNet. In Intelligent Information Systems XVI, Proceedings of the International IIS"08 Conference. Warszawa : Academic Publishing House EXIT, 2008. od s. 369-374, 6 s. ISBN 978-83-60434-44-4.

Horák, Aleš - Vossen, Piek - Rambousek, Adam. The Development of a Complex-Structured Lexicon based on WordNet. In Proceedings of the Fourth Global WordNet Conference. Szeged : University of Szeged, 2008. od s. 200-208, 9 s. ISBN 978-963-482-854-9.

Horák, Aleš - Pala, Karel - Rambousek, Adam. The Global WordNet Grid Software Design. In Proceedings of the Fourth Global WordNet Conference. Szeged : University of Szeged, 2008. od s. 194-199, 6 s. ISBN 978-963-482-854-9.

Horák, Aleš - Pala, Karel - Rambousek, Adam. Tools for Managing Multilingual Lexical Resources. In Procedings of the 16th International Conference Intelligent Information Systems. Zakopane, Poland : Polish Academy of Sciences, 2008. od s. 451-460, 10 s. ISBN 978-83-60434-44-4.

Číslo aktivity
26/08
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Tvorba modelu české syntaxe na základě korpusu syntaktických stromů
Zahájení aktivity
1.1.2008
Ukončení aktivity
31.12.2008
Popis aktivity
Korpus syntaktických stromů byl rozšířen na více než 6000 vět s vyznačenou složkovou syntaktickou strukturou. Tento korpus se aktivně využívá při vývoji a testování algoritmů na modelování české syntaxe.
Skutečné Indikátory dosažení - výsledky aktivity
zkvalitněný model české syntaxe použitý v syntaktickém analyzátoru synt.
Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity
rozšířený korpus syntaktických stromů pro 6163 vět.

publikace:
- Horák, Aleš. Computer Processing of Czech Syntax and Semantics. 1st edition. Brno, Czech Republic : Librix.eu, 2008. 241 s. 1st edition. ISBN 978-80-7399-375-7

- Jakubíček, Miloš. Extraction of Syntactic Structures Based on the Czech Parser Synt. In Proceedings of Recent Advances in Slavonic Natural Language Processing 2008. Brno : Masaryk University, 2008. s. 56-62. ISBN 978-80-210-4741-9.

- Kovář, Vojtěch - Jakubíček, Miloš. Test Suite for the Czech Parser Synt. In Proceedings of Recent Advances in Slavonic Natural Language Processing 2008. Brno : Masaryk University, 2008. s. 63-70. ISBN 978-80-210-4741-9.

Číslo aktivity
27/08
Ke kterému dílčímu cíli se aktivita vztahuje
2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...
Název (cíl)aktivity
Návrh formalismu pro práci s konstrukcemi TILu jako sémantické reprezentace českých vět
Zahájení aktivity

1.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Pro zkvalitnění logické analýzy věty v přirozeném jazyce pokračovaly práce na podrobném návrhu formalismu pro tvorbu konstrukcí a typování vstupních slov. Formalismus pro tvorbu konstrukcí je založen (v souladu s principem kompozicionality) na pravidlech syntaktického analyzátoru. Typování vstupních slov sleduje zvolenou ontologii - hypero/hyponymickou hierarchii z princetonského WordNetu a významy a struktura predikátů využívají budované databáze VerbaLex.

Skutečné Indikátory dosažení - výsledky aktivity

Rozšířený popis formalismu tvorby konstrukcí TIL.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

- Horák, Aleš. Computer Processing of Czech Syntax and Semantics. 1st edition. Brno, Czech Republic : Librix.eu, 2008. 241 s. 1st edition. ISBN 978-80-7399-375-7.

- Pala, Karel - Horák, Aleš. Can Complex Valency Frames be Universal? In RASLAN 2008. 1. vyd. Brno : Masarykova Univerzita, 2008. od s. 41-48, 8 s. ISBN 978-80-210-4741-9

Číslo aktivity

28/08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh a implementace guesseru - modulu pro automatické doplňování morfologické databáze češtiny

Zahájení aktivity

1.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

V rámci aktivity vznikl algoritmus a zkušební implementace guesseru. Guesser umožňuje rozšiřovat a doplňovat morfologickou databázi češtiny zejména s ohledem na přistupování k webu a zpracování termínů.

Skutečné Indikátory dosažení - výsledky aktivity

základní verze guesseru

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

- Šmerk, Pavel. Towards Czech Morphological Guesser. In Sojka, Petr - Horák, Aleš. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008. Brno : Masarykova univerzita, 2008. od s. 1-4, 4 s. ISBN 978-80-210-4741-9

- Pala, Karel - Svoboda, Lukáš - Šmerk, Pavel. Czech MWE Database. In Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 08). Marrakech, Morocco : European Language Resources Association (ELRA), 2008. s. 1-5. ISBN 2-9517408-4-0.

Číslo aktivity

29/08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Detekce plagiátů (spamů) s využitím sémantických znalostí

Zahájení aktivity

1.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Zjištění aplikovatelnosti vektorových modelů pro určení textové podobnosti dokumentů. Dokumenty pocházejí z digitálních knihoven. Jsou nestrukturované a často jsou výsledkem OCR, tj. chyby vznikají už na písmenové úrovni. Proběhlo srovnání různých metod z oblasti Information Retrieval, které se snaží určit dokumentovou podobnost za vynaložení přijatelných výpočetních prostředků i pro velké kolekce dat (desítky tisíc dokumentů).

Skutečné Indikátory dosažení - výsledky aktivity

Jako prostředek pro dosažení těchto cílů vznikla řada skriptů, které proces výroby podobnostních matic a hledání plagiátů automatizují. Pro SVD byla využita knihovna PROPACK. Skripty běží automatizovaně v projektu DML-CZ. Je plánováno zobrazení textové podobnosti dokumentů uživateli přes webové rozhraní.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Řehůřek, Radim. Plagiarism Detection through Vector Space Models Applied to a Digital Library. In *RASLAN 2008*. 1,. Brno : Masarykova univerzita, 2008. od s. 75-83, 9 s. ISBN 978-80-210-4741-9.

Číslo aktivity

30/08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Rozpoznávání anaforických vztahů ve volných textech

Zahájení aktivity

1.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

V rámci aktivity probíhaly práce na návrhu a implementaci programu, který automaticky vyhledává a analyzuje anaforické vztahy ve volných textech.

Skutečné Indikátory dosažení - výsledky aktivity

Řešení otázek spojených s použitím algoritmů pro řešení anaforických vztahů (AR) na výstup syntaktického analyzátoru synt. Zobecňování a optimalizace stávající implementace systému na řešení anaforických vztahů. Příprava začlenění dalších zdrojů dat použitelných při AR - zejména český WordNet, valenční slovník VerbaLex a statistiky Sketch Engine.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

Němčík, Václav. The Saara Framework: Work in Progress. In: RASLAN 2008, 2. vyd. Brno : Masarykova Univerzita, 2008. od s. 11-16, 6 s. ISBN 978-80-210-4741-9.

Číslo aktivity

31/08

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh a vývoj algoritmů pro vytváření grafiky a webovských prezentací prostřednictvím dialogových systémů.

Zahájení aktivity

1.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

V průběhu dílčího období pokračovaly práce na technologiích souvisejících se zpřístupňováním grafických objektů pro zrakově postižené uživatele. Byla analyzována struktura grafických ontologií a vytvořen prototyp grafické ontologie pro použití při anotaci grafických objektů. Byl vytvořen a implementován algoritmus pro konverzi formátu JPEG do formátu SVG integrujícího bitmapovou strukturu přímo ve vlastním formátu, což zajišťuje bezproblémovou přenositelnost bitmapových grafických objektů prostřednictvím formátu SVG. Byla vytvořena koncepce a struktura základních modulů pro anotátor grafických objektů ve formátu SVG. Byla testována metoda strukturálního popisu grafických scén. Pokračovala práce na metodách integrujících popis grafického objektu do formátu SVG a byla testována technologie využití tohoto přístupu pro nevidomé uživatele. Pokračovalo testování systému WebGen pro vytváření webovských presentací dialogovým způsobem s cílem zefektivnění použitých dialogových strategií. Do systému byla přidána podpora pro nové sémantické typy presentací. Dále bylo úspěšně otestováno vygenerování jednotlivých stránek. a byly rovněž provedeny testy nevidomými uživateli včetně ověřeni požadavku přístupnosti (internetový standard Web Content Accessibility). Byla vytvořena první verze systému WebGen.

Skutečné Indikátory dosažení - výsledky aktivity

Publikace:

Bártek, L., Plhák, J. - Visually Impaired Ureser Create Web Pages, in
In Computers Helping People with Special Needs: 11th International Conference, ICCHP 2008. Berlin : Springer-Verlag, 2008. od s. 466-473, ISBN 3-540-70539-2

Kopeček, Ivan - Ošlejšek, Radek. GATE to Accessibility of Computer Graphics.
In Computers Helping People with Special Needs: 11th International Conference, ICCHP 2008. Berlin : Springer-Verlag, 2008. od s. 295-302, ISBN 3-540-70539-2

Kopeček, Ivan - Ošlejšek, Radek. Dialogue-Based Processing of Graphics and Graphical Ontologies. 11th International Conference, TSD 2008. Berlin : Springer-Verlag, Brno 2008.

Implementace:
První verze systému WegGen
Algoritmus pro konverzi formátu JPEG do formátu SVG integrujícího bitmapovou strukturu přímo ve formátu SVG.
Prototypová verze grafické ontologie.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

viz příslušné sborníky

Číslo aktivity

32/08

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Klasifikace matematických textů ve vytvořeném korpusu

Zahájení aktivity

1.1.2008

Ukončení aktivity

31.12.2008

Popis aktivity

Byly a zpracována rešerše používaných klasifikačních schémat pro matematické texty a analyzovány možnosti využití klasifikace pro zpracování (desambiguace) a vyhledávání v korpusu matematických textů. Byl shromážděn korpus téměř 200,000 stran matematických textů (knihovny DML-CZ, NUMDAM, arXiv) a jeho část použita pro vytvoření klasifikátoru a návrhu měření podobnosti matematických článků.

Skutečné Indikátory dosažení - výsledky aktivity

Vyvinutý klasifikátor je použitelný pro klasifikaci hlavní úrovně Mathematical Subject Classification u retro-digitalizovaných článků, kdy MSC ještě nebyla rozšířena. Je testován a vyhodnocován v projektu DML-CZ spolu s podobnostní maticí spočtenou pro shromážděné články třemi metodami.

Skutečné prostředky ověření - forma zpracování a předání výsledku aktivity

SOJKA, Petr - ŘEHŮŘEK, Radim. Automated Classification and Categorization of Mathematical Knowledge. In: Intelligent Computer Mathematics: AISC/Calculemus/MKM LNAI 5144. Vyd. první. Berlin, Heidelberg, New York : Springer-Verlag, 2008. ISBN 978-3-54085109-7, s. 543-557. 28.7.2008, Birmingham.

SOJKA, Petr. DML 2008 Towards Digital Mathematics Library. Brno, Czech Republic: Masaryk University Press, 2008. 183 s. mimo edice. ISBN 978-80-210-4658-0

SOJKA, Petr. Towards Natural Natural Language Processing. In: RASLAN 2008 Proceedings. Vyd. prvni. Brno : Masaryk University, 2008. ISBN 978-80-210-4741-9, s. 98-100. 5.12.2008, Karlova Studánka.

SOJKA, Petr - HORÁK, Aleš. Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008. Edited by Sojka P., Horák A. Vyd. prvni. Brno: Masaryk University, 2008. 110 s. RASLAN Proceedings. ISBN 978-80-210-4741-9.

2.2.2. AKTIVITY NEUSKUTEČNĚNÉ v roce 2008

Číslo aktivity
Ke kterému dílčímu cíli se aktivita vztahuje
Název (cíl)aktivity
Zahájení aktivity
Ukončení aktivity
Popis aktivity
Důvody, proč se aktivitu nepodařilo uskutečnit

2.3.NÁKLADY PROJEKTU - 2008

2.3.1. NÁKLADOVÉ TABULKY ZA JEDNOTLIVÉ SUBJEKTY

Rok	2008
Typ	skutečné
Organizace	Západočeská univerzita v Plzni
Role organizace	příjemce - koordinátor

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**2965**	**2945**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**0**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**100**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**70**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**130**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**380**	**50**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**350**	**50**
F9. CELKEM	**3995**	**2995**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**75**	**75**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**3995**	**2995**	**0**	**1000**

Rok	2008
Typ	skutečné
Organizace	Masarykova univerzita
Role organizace	spolupříjemce

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**1822**	**1521**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**60**	**40**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**55**	**42**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**0**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**0**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**185**	**119**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**200**	**0**
F9. CELKEM	**2322**	**1722**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**2322**	**1722**	**0**	**600**

2.3.2. NÁKLADOVÁ TABULKA ZA PROJEKT

Rok	2008
Typ	skutečné
PROJEKT	2C06009 - CELKEM

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady skutečně vynaložené tis. Kč	z toho skutečně hrazené z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**4787**	**4466**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**60**	**40**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**155**	**42**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**70**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**130**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**565**	**169**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**550**	**50**
F9. CELKEM	**6317**	**4717**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**75**	**75**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**6317**	**4717**	**0**	**1600**

2.3.3. ZDŮVODNĚNÍ ZMĚN V ČERPÁNÍ

Dopisem z 15.10.2008 bylo MŠMT požádáno o povolení změny položkového členění uznaných nákladů projektu 2C06009 v roce 2008. Změna (přesun) se týkala pouze prostředků ze spoluúčasti řešitelského pracoviště ZČU a neměnila celkově uznané náklady. Důvodem přesunu mezi položkami byla zejména nemožnost účtovat náklady na ubytování při zahraničních služebních cestách do položky „cestovní náklady“. Požádali jsme proto o svolení s následující úpravou:
Snížení položky F7 „Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu“ o 70.000,-Kč.
Navýšení položky F5 „Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu“ o 40.000,-Kč.
Navýšení položky F6 „Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu“ o 30.000,- Kč.

Tuto změnu MŠMT dopisem ze dne 11.11.2008 schválilo.

Centrum ZPJ FI MU čerpalo prostředky projektu podle původního plánu, pouze částka určená na cestovné byla o 5.000,-Kč navýšena, přičemž o stejnou částku byly nižší provozní náklady. Souhrnné náklady za projekt tedy zůstaly beze změn.

2.3.4. NEVYUŽITÉ FINANČNÍ PROSTŘEDKY

Veškeré poskytnuté finanční prostředky byly pro řešení projektu využity.

2.3.5. Seznam hmotného a nehmotného majetku pořízeného za sledované období

3. ZÁMĚR A NÁVRHY PRO NÁSLEDUJÍCÍ OBDOBÍ - rok 2009

3.1. PROJEKTOVÝ TÝM A ŘEŠITELSKÉ TÝMY

3.1.1. PROJEKTOVÝ TÝM

IČ organizace	49777513
Obchodní jméno - název	**Západočeská univerzita v Plzni**
Zkratka názvu	ZČU
Role organizace	příjemce - koordinátor
Vazba na organizaci	00216224
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Univerzitní 8/
- PSČ, obec	30614 Plzeň
- stát	Česká republika
- telefon	377 631 111
- http://	www.zcu.cz
Bankovní spojení	
-DIČ	CZ49777513
- banka kód, název	0100 - Komerční banka, a.s., Plzeň
- číslo účtu, sp.symbol	4811530257,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Doc. Ing. Josef Průša CSc.
- funkce	rektor
- telefon	377631000
- mobil	606665105
- fax	377631002
- email	rektor@rek.zcu.cz

IČ organizace	00216224
Obchodní jméno - název	**Masarykova univerzita**
Zkratka názvu	MU
Role organizace	spolupříjemce
Vazba na organizaci	49777513
Druh organizace	Veřejná nebo státní vysoká škola (zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (o vysokých školách)
Adresa sídla, spojení na organizaci	
- ulice, čp./č.or.	Žerotínovo náměstí 617/ 9
- PSČ, obec	60177 Brno
- stát	Česká republika
- telefon	549 491 1111
- http://	www.muni.cz
Bankovní spojení	
-DIČ	CZ00216224
- banka kód, název	0100 - Komerční banka Brno-město
- číslo účtu, sp.symbol	85636621,
Statutární zástupce	
- titul před, jméno, příjmení, titul za	Prof. PhDr Petr Fiala PhD
- funkce	rektor
- telefon	549491001
- mobil	
- fax	
- email	rektor@muni.cz

3.1.2. ŘEŠITELSKÝ TÝM

Celé jméno, RČ	**Bártek Luděk Mgr.** 7201083791 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3215 bar@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Brada Přemysl Ing. PhD. MSc.** 7007012111 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	3772435 brada@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Češka Zdeněk Ing.** 8207311244 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632452 zceska@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Ekštein Kamil Ing. PhD.** 7705302011 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 kekstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Habernal Ivan Ing.** 830705/1764 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 habernal@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Hejtmánek Jan Ing.** 821101/2095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 377 632 402 hejtman2@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	10

Celé jméno, RČ	**Horák Aleš RNDr. Ph.D.** 7409014250 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 4377 hales@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Hynek Jiří ing. PhD.** 720506/2029 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632455 hynekj@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Ježek Karel doc. Ing. CSc.** 420617110 CZ
Role osoby při řešení projektu	řešitel
Spojení	377 632 475 jezek_ka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Klečková Jana doc. Dr. Ing.** 496108095 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 421 kleckova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Konopík Miloslav Ing.** 8103261782 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 konopik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Kopeček Ivan doc. RNDr. CSc.** 490303075 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 3861 kopecek@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Král Pavel Ing. PhD.** 760317/2049 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632454 pkral@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Krutišová Jana Ing.** 5955160046 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 413 krutisova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

Celé jméno, RČ	**Matoušek Václav prof. Ing. CSc.** 480613108 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 471 matousek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mautner Pavel Ing. PhD.** 6505222592 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 mautner@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Mouček Roman Ing. PhD.** 7607072000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 441 moucek@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Pala Karel doc. PhDr. CSc.** 390615416 CZ
Role osoby při řešení projektu	spoluřešitel
Spojení	549 49 5616 pala@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Pavelka Tomáš Ing.** 7909182083 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Pomikálek Jan Mgr.** 7910090419 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 1864 xpomikal@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	

Celé jméno, RČ	**Ptáčková Helena** 705914/2079 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 463 377 632 402 ptackova@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Rambousek Adam Bc.** 811022/5233 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	xrambous@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	

Celé jméno, RČ	**Rohlík Ondřej Ing. PhD.** 7510031925 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632450 rohlik@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	50

Celé jméno, RČ	**Rychlý Pavel Mgr. Ph.D.** 7301235359 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549 49 6399 pary@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Sojka Petr doc. RNDr. Ph.D.** 6309171000 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	549496966 sojka@fi.muni.cz
Příslušnost k organizaci	Masarykova univerzita Fakulta informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	

Celé jméno, RČ	**Steinberger Josef Ing. PhD.** 7909182127 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377 632 479 jstein@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	20

Celé jméno, RČ	**Toman Michal Ing.** 8007042054 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632479 mtoman@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	pracovník přijatý na dobu řešení projektu
Pracovní kapacita v %	100

Celé jméno, RČ	**Zíma Martin Ing. PhD.** 7405042073 CZ
Role osoby při řešení projektu	člen řešitelského týmu
Spojení	377632431 zima@kiv.zcu.cz
Příslušnost k organizaci	Západočeská univerzita v Plzni Katedra informatiky
Pracovní poměr	kmenový pracovník organizace
Pracovní kapacita v %	10

3.1.3. ZMĚNY V PROJEKTOVÉM A ŘEŠITELSKÝCH TÝMECH - rok 2009

Pč.	Typ	Popis
1	návrhy změn v projektovém týmu a řešitelských týmech	K 31.12.2008 práci v projektovém týmu Centra ZPJ FI MU ukončil Patrick Hanks a jako nový člen týmu byl přibrán Adam Rambousek.

3.2. ČASOVÝ POSTUP PRACÍ - rok 2009

3.2.0. PŘEHLED DÍLČÍCH CÍLŮ PLÁNOVANÉ 2009

	Číslo	Dílčí cíl podrobně	Datum plnění
	1	**Dílčí cíl** Příprava a vytvoření datových kolekcí (soubory, korpusy) pro trénování a ověřování algoritmů komunikace s www prostředím. **Indikátory dosažení - výsledky dílčího cíle** a) Vytvoření uživatelského rozhraní pro hlasový vstup / příp. výstup, které bude použito pro komunikaci se sémantickým webem, a pro jeho podporu vytvoření robustního ASR systému pro inflexní jazyky. K tomu bude nutno vytvořit kvalitní korpus pro ASR a z něj extrahovat dostatečné množství trénovacích dat. V jednotlivých etapách bude v průběhu let 2006 – 2007 vytvořen: - kvalitní audio-korpus pro natrénování systému ASR, - korpus pro natrénování jazykových modelů. b) Příprava datových kolekcí a pomocných rutin vyhledávacího systému ve vícejazyčných korpusech, včetně prostředků pro zpřesňování uživatelských dotazů pomocí thesauru a nástrojů pro disambiguaci víceznačných slov, na bázi klient/server aplikace. Jednotlivé dílčí výsledky řešení projektu lze charakterizovat takto: - vytvoření multijazykových korpusů – základní výběr zahrnuje angličtinu a češtinu, dle možností alespoň některé úlohy plánujeme provádět i se slovenštinou (zajímavá je blízkostí k češtině) a němčinou, - metoda automatického rozpoznání jazyka – kombinace „stop slov“ a frekvenčních znakových metod. c) Příprava datových kolekcí a modulů pro filtraci a sumarizaci textů: - vytvoření sumarizačních korpusů (pro angličtinu plánujeme využít standardních korpusů, např. DUC a pro češtinu bude vytvořen vlastní, složený vesměs z textů novinových článků, - sumarizace textů založená na latentní sémantické analýze (LSA), vytvoření anotované kolekce pro sumarizátor založený na LSA - vytvoření vícejazyčných korpusů, - rozšíření standardních textových korpusů o korpusy závadných dokumentů pokrývající problematická témata definovaná v zadání. d) Korpus syntaktických stromů (treebank): - korpus bude morfologicky označkován a zjednoznačněn, - bude v něm vyznačena závislostní struktura věty i jednotlivé větné složky včetně koreferenčními vztahy, - korpus bude z části založen na existujícím PDT. e) Korpus vzorových přepisů vybraných vět a jejich sémantické reprezentace: - text korpusu bude podmnožinou korpusu syntaktických stromů, - ve stromech budou vyznačeny významy z dostupných ontologií (WordNet), - věty budou rozšířeny o logické formy. f) Doplnění morfologického značkovače o robustní hádací proceduru, která bude spolehlivě přiřazovat morfologické značky i neznámým slovům. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Jedná se o vytvoření podpůrného aparátu, bez něhož nelze další zamýšlené cíle projektu dosáhnout. Vytvořeny budou proto korpusy v podobě rozsáhlých datových souborů se specifickou strukturou a organizací a pro jejich údržbu a prohledávání budou vyvinuty speciální softwarové nástroje. Výsledky budou soustředěny do soustavy datových souborů a její obsah prezentován formou publikace na konferencích a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující náplň dílčího cíle „1“ a nástin jejich řešení jsou následující: RF1: Během zpracování korpusů a korpusových nástrojů se vyskytnou další korpusy obsahující srovnatelná data. Řešení: Korpusy pro český jazyk vznikají v ČR na celkem pěti pracovištích, která udržují těsné kontakty a výsledky výzkumu si vzájemně vyměňují nebo se o nich poměrně obsáhle	- 31.12.2007

		informují. Navíc je třeba rozlišovat mezi korpusy psanými (textovými) a řečovými. Řečové korpusy vznikají prakticky jen na pracovištích v Plzni, Brně a Liberci, z nichž dvě se na řešení tohoto projektu budou podílet. Navíc vznik jakéhokoli dalšího korpusu je pozitivním jevem, neboť v tomto oboru více než kdekoli jinde platí, že vhodných dat není nikdy dostatek. Tudíž korpusy vytvořené v rámci navrhovaného projektu budou v každém případě využity i dalšími pracovišti. V případě cizojazyčných korpusů budou využívány korpusy, které jsou k dispozici v systému ELRA (European Language Resources Association). RF2: Nepodaří se získat dostatek materiálů, resp. mluvčích, pro vytvoření textových, resp. audiokorpusů. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť již současný web poskytuje doslova nepřeberné množství textového materiálu, z nichž lze za použití vhodných vyhledávacích metod vybrat dostatečné množství materiálu pro vytvoření korpusu. V případě řečových korpusů nejde ani tak o problém nalezení vhodné množiny dat nebo množiny vhodných mluvčích, nýbrž kritickým faktorem je čas. Pořizování řečových dat a zejména jejich následné zpracování (třídění, anotace, apod.) vyžaduje značné množství času, avšak riziko lze úspěšně odstranit kvalitním managementem projektu. RF3: V průběhu naplňování dílčího cíle projektu se vyskytne komerční software řešící problematiku pořizování korpusů. Řešení: Pokud se nějaký software vyskytne a bude využitelný, nebude díky modularitě předpokládaného programového vybavení příliš obtížné ho do vytvářeného software začlenit. Pravděpodobnost jeho výskytu v dohledné době je však minimální.	
	2	**Dílčí cíl** Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka. **Indikátory dosažení - výsledky dílčího cíle** a) Návrh formalismu pro popis sémantiky na rozsáhlejší doméně, návrh vhodně strukturovaného sémantického popisu dotazů uživatelů, eventuálně vytvoření vlastního hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému. b) Vytvoření ontologií pro aplikaci formalismu popisujícího sémantiku. Jednotlivými výsledky budou: - návrh ontologie, sémantických konceptů – datový formát XML, vytvoření UML modelu, - návrh ohodnocení jednotlivých konceptů vektorem sémantických příznaků, a to jak doménových, tak obecnějšího charakteru, - návrh soustavy vektorů ohodnocení jednotlivých konceptů. c) Vytvoření multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí, jeho zakomponování do prostředí pro vyhledávání a vývoj metod ohodnocování jeho kvality, návrh metod disambiguace v multijazykovém prostředí s využitím kontextu, thesauru a pravděpodobnostních metod: - sumarizační systém obohacený o kompresi souvětí, - systém rezoluce anafor a jeho využití při sumarizaci – pro angličtinu bude využit systém GuiTAR, vytvořený na univerzitě Essex (Anglie), pro češtinu bude na základě poznatků získaných na českých pracovištích vytvořen vlastní systém, - metoda hodnocení kvality sumarizátorů na základě LSA. d) Vývoj nových, dokonalejších modelů elektronických dokumentů tak, aby při použití textových klasifikačních algoritmů bylo dosaženo co nejlepších výsledků při rozpoznávání tématu, rozpoznávání spamových emailů, detekci dokumentů se závadným obsahem apod. e) Vytvoření metodologie a nástrojů pro analýzu webových dokumentů. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Při naplňování tohoto dílčího cíle půjde o vytvoření základního teoretického podpůrného aparátu, bez něhož nebude možné další kroky realizovat. Jediný tento dílčí cíl bude mít charakter spíše základního výzkumu – půjde o vývoj metod, metodologií a formálních modelů pro návrh zamýšleného komunikačního rozhraní, avšak součástí výzkumných prací bude též experimentální implementace a vytvoření softwarových nástrojů pro evaluaci vyvíjených metod a formalismů. Výsledky budou shrnuty do písemných dokumentů a prezentovány téměř výhradně formou publikací na konferencích, v odborných časopisech a v průběžných výzkumných zprávách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „2“ a nástin jejich řešení mohou být následující:	- 31.12.2008

		RF1: Nepotvrzení či neplatnost výzkumných hypotéz poskytujících základ pro vytvoření formalismů a modelů. Řešení: Plánovaný dílčí cíl zde nestojí na jediné výzkumné hypotéze, nýbrž na teoretickém základu návrhu komunikačních systémů. Využito bude jak dosavadních poznatků z návrhu existujících komunikačních rozhraní a systémů pro interakci člověka s počítačem, tak i poznatků z psychologie komunikace a doporučení TC.13 IFIP (for HCI). Základním rizikem proto bude opět časový faktor, který lze výrazně omezit dobrým managementem projektu. RF2: Nedostatečná erudice členů týmu pro vývoj formálních prostředků. Řešení: Tento rizikový faktor nebude mít zřejmě přílišnou váhu, neboť oba participující týmy jsou složeny minimálně z poloviny ze starších zkušených výzkumníků, z nichž někteří se předmětnou oblastí zabývají 25 i více let, z druhé části pak z mladých perspektivních pracovníků, kteří buď vyrostli anebo se podíleli na řešení podobné problematiky a potřebné teoretické základy oboru již získali, zejména v doktorandském studiu.	
	3	**Dílčí cíl** Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. **Indikátory dosažení - výsledky dílčího cíle** a) Implementace uživatelského rozhraní pro hlasovou komunikaci se sémantickým webem –součástí výsledku budou: - implementace LVCSR rozpoznávače, - natrénování akustických a jazykových modelů, - implementace nahrávacího modulu se stochastickým modelem detekce řečového signálu, - implemetace parametrizátoru na bázi MFCC, - návrh a implementace modulu pro akustické modelování založeného na umělých neuronových sítích nebo směsích Gaussových funkcí, - návrh a implementace efektivního dekódovacího algoritmu, který dokáže pracovat s gramatikami a stochastickými jazykovými modely, - programová realizace a ověření funkčních vlastností robustního ASR systému pro inflexní jazyky. b) Systém pro extrakci významu ze spontánních promluv – dílčími kroky k dosažení tohoto dílčího cíle budou: - návrh a realizace optimální řečové databáze, - návrh systému sémantického značkování řečových dat, - báze znalostí umožňující automatizované či automatické značkování spontánních promluv uložených v databázi, - implementace stochastických sémantických gramatik pro automatickou sémantickou analýzu dotazu uživatele, - využití hierarchické ontologie pro tvorbu strukturalizovaného popisu dotazů uživatele a pro zajištění schopnosti zobecňování z natrénovaných dat, - aplikace metod mělkého (shallow) parsingu promluv pro částečnou analýzu dotazů uživatele. c) Vytvoření komfortního uživatelského rozhraní pro práci se sémantickým webem – součástí tohoto dílčího cíle bude: - návrh příslušného dialogového manageru akceptujícího tzv. kombinovanou iniciativu ve vedení dialogu (mixed initiative), - vytvoření robustního systému pro efektivní a časově nenáročné vyhledávání dat v řečové databázi, - vytvoření robustního a spolehlivého modelu sémantické hierarchie a jeho implementace. d) Aplikace a modifikace OWL standardu v českém prostředí. e) Aplikace klasifikačních metod v multijazykovém prostředí. f) Kompletace multilingválního sumarizačního systému včetně rezoluce anafor a komprese souvětí. g) Algoritmy vhodné pro generování itemsetů a n-gramů a ověření jejich úspěšnosti pro klasifikaci textových dokumentů. h) Výchozí algoritmy pro vyvozování nových znalostí z informací získaných z volného textu. i) Prototyp programu pro přiřazování logických formulí větám z volného textu. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** V dílčím cíli „3“ jde o vytvoření souboru programových produktů, které vzniknou implementací teoretických metod a formalismů vytvořených v rámci dílčího cíle „2“. Výsledky budou mít jednoznačně aplikační charakter, i když vesměs půjde jen o experimentální software, bez	- 31.12.2009

		něhož nelze metody a modely verifikovat. Výsledky však bude možno předat i dalším zájemcům, protože se předpokládá úplná dokumentace vytvořeného programového vybavení. Výsledky budou prezentovány jako balíky experimentálního software a metod, dále budou publikovány na konferencích, v průběžných výzkumných zprávách, případně také zveřejněny formou speciálních letáků, v tisku a uvažuje se též o možnosti předvedení na specializovaných veletrzích a výstavách. **Kritické poedpoklady dosažení dílčího cíle** Rizikové faktory ovlivňující dosažení dílčího cíle „3“ a nástin jejich řešení: RF1: V průběhu projektu přestane být o vytvářené přístupové technologie zájem a pracoviště účastnící se na řešení projektu se tak ocitnou bez reálné využitelnosti svých výsledků. Řešení: Současným trendem je naopak příklon k využívání multimediálních a multimodálních dat, ukládání velkých množství dat a informací na běžných počítačových prostředcích, sílí propojování informačních technologií s rozhlasovým a televizním vysíláním, streamovanými médii a mobilními komunikacemi. Nové hardwarové prostředky budou vyžadovat nové technologie přístupu k datům, přičemž preferována bude komunikace v přirozeném jazyce, ať už psanou nebo mluvenou formou. Vyvíjené programové prostředky tento trend jednoznačně podpoří a proto je toto riziko za dobu řešení projektu téměř nulové. RF2: V průběhu řešení projektu se vyskytne komerční software řešící problematiku srovnatelnou s předpokládanými výsledky projektu. Řešení: Komerční řešení využívající přístup k datům na webu prostřednictvím přirozeného jazyka jsou dosud v plenkách a komerční sféra naopak aktivně vyhledává zajímavé práce z akademické sféry. Proto je toto riziko minimální, očekáváme naopak velký zájem z komerční sféry. RF3: Časové faktory ovlivňující zpracování software. Řešení: Při implementaci a programové realizaci metod vyvinutých v rámci dílčího cíle „2“ může dojít k určité časové tísni vlivem nevhodně zvolených implementačních nástrojů, eventuálně nezkušeností některých mladších členů týmu. Riziko je však minimální, neboť řešitelský kolektiv je složen vesměs ze zkušených výzkumníků a mladých pracovníků, kteří již obdobné, i když jednodušší systémy v minulosti vytvářeli a implementovali. Časový faktor lze navíc výrazně ovlivnit dobrým managementem projektu.	
	4	**Dílčí cíl** Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí. **Indikátory dosažení - výsledky dílčího cíle** a) Testování a ověřovací provoz implementovaného hlasového rozhraní – součástí bude - otestování zpracovaného LVCSR rozpoznávacího systému, - ověření funkčních vlastností robustního ASR systému na vhodné množině uživatelů, - otestování vyvinutých metod automatické sémantické analýzy dotazů. b) Ověření funkčních vlastností vytvořených ontologií a hierarchického systému relací mezi lexémy pro zaručení generalizační schopnosti systému analýzy sémantiky, c) Ověření vlastností algoritmů pro klasifikaci a analýzu dat na různých typech dokumentů. d) Otestování a ověření navržených metod na konkrétních typových řešeních, např. na přístupu k webovým stránkám výzkumných a vzdělávacích institucí. e) Vyhodnocení úspěšnosti jednotlivých fází analýzy volného textu od morfologické úrovně až po převod do logických formulí. **Prostředky ověření - Forma zpracování a předání výsledku dílčího cíle** Náplní dílčího cíle „3“ je provedení rozsáhlých testů (tzv. field experiments) vyvinutých metod, metodologií, modelů a vytvořeného souboru programových produktů. Předpokládá se testování produktů na obvyklých třech skupinách uživatelů – v prvním kroku budou vlastnosti systémů a metod prověřovány úzkou skupinkou řešitelů projektu, ve druhém kroku bude testovací množina uživatelů vytvořena ze spolupracovníků, kteří však s řešením projektu neměli nic společného a o výsledcích řešení jsou jen velmi kuse informováni, a teprve ve třetím kroku bude systém testován libovolnými uživateli, tzv. „lidmi z ulice“. Zčásti však v tomto kroku budou využiti studenti, kteří všeobecně mají tendenci takové systémy „pokořit“. Výsledky budou kompletně dokumentovány a z vyhodnocení experimentů budou vyvozovány příslušné závěry, tj. systém a jeho části budou průběžně doplňovány, upravovány a opětovně testovány. V závěru budou výsledky testování a ověřovacího provozu publikovány v časopisech, na konferencích a obšírně v závěrečné výzkumné zprávě.	- 31.12.2010

Kritické poedpoklady dosažení dílčího cíle
Rizikové faktory ovlivňující dosažení dílčího cíle „4“ a možná řešení:

RF1: V průběhu testů se projeví nedostatky v koncepci systému vedoucí k závažným problémům ve funkci systému.
Řešení: Řešitelský tým je složen z odborníků, kteří obdobné, i když jednodušší, systémy již vytvořili a mají z jejich tvorby nezanedbatelné zkušenosti. Tým byl dále doplněn o mladé pracovníky, kteří se podíleli na tvorbě řady produktů pro prezentace na webových stránkách a je jim problematika přístupu k webu velmi blízká. Riziko volby nevhodné koncepce je proto minimální.

RF2: V průběhu testů se projeví nedostatky v implementaci systému a metod.
Řešení: Obdobné jako předchozí rizikový faktor – řešitelský tým je složen z odborníků, kteří obdobné, systémy již vytvořili a mají i z jejich implementace poměrně rozsáhlé zkušenosti. Riziko závažných implementačních chyb je proto minimální, drobné nedostatky v implementaci bývají zpravidla v krátké době snadno odstranitelné.

RF3: Nepodaří se vytvořit dostatečně reprezentativní množiny testovacích osob.
Řešení: Ve vztahu k odstavci 3.3.3. (tři úrovně testování) je riziko nedostatečného vytvoření skupin testujících osob nepatrné – obě participující pracoviště jsou poměrně rozsáhlá a množinu osob testujících vlastnosti systému nebude problém vytvořit; ostatně bylo již ověřeno v minulosti na jednodušších úlohách. Otázka volby třetí skupiny osob je spíše otázkou vytvořeného přístupu k systému – zde se nabízejí dvě možnosti: Buď si osoby vhodné k testování systému vybírat podle určitých hledisek (bylo tak někdy postupováno v minulosti a osoby byly k testování zvány na řešitelské pracoviště) nebo zveřejnit přístupový portál systému a dovolit testování systému široké veřejnosti prostřednictvím internetu, popř. přes telefon (telefonní přístup je však v současných podmínkách omezen kvalitou spojení v mobilních sítích, resp. kvalita spojení je dána úrovní signálu v místech, kde se potenciální uživatel právě nachází, a výsledky testů jím mohou být zkresleny). Rizikový faktor může být opět minimalizován vhodnými rozhodnutími, resp. dobrým managementem projektu.

3.2.1. AKTIVITY PLÁNOVANÉ NA DALŠÍ OBDOBÍ - rok 2009

Číslo aktivity
01/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Vývoj rozpoznávače JLASER
Zahájení aktivity
5.1.2009
Ukončení aktivity
22.12.2009
Popis aktivity
Automatický rozpoznávač řeči JLASER je nyní ve verzi 1.2, předpokládá se, že vývoj bude nadále pokračovat podle potřeb projektu. Jako perspektivní se v současné době jeví implementace PLP parametrizátoru, který pracuje lépe ve zhoršených akustických podmínkách. Rovněž se předpokládá další práce na dekodéru ve spojení se stochastickými jazykovými modely.
Plánované indikátory dosažení - očekávané výsledky aktivity
Bude provedeno rozsáhlé testování zpracovaného programového systému a vyhodnoceny testy úspěšnosti rozpoznávání.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Zdrojové kódy programového řešení budou umístěny na webovou stránku projektu.

Číslo aktivity
02/09
Ke kterému dílčímu cíli se aktivita vztahuje
4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....
Název (cíl)aktivity
Poměrové statistiky
Zahájení aktivity
5.1.2009
Ukončení aktivity
22.12.2009
Popis aktivity
Při testování úspěšnosti rozpoznávání je často problém nedostatku dat. Pokud jsou např. výsledky dvou testů blízko, musí se vzít v úvahu náhodná chyba a je třeba testovat statistickou významnost. Existující statistické testy většinou předpokládají nezávislost sledovaných jevů (v našem případě je sledovaným jevem rozpoznání slova). Při rozpoznávání vět může chyba v jednom slově způsobit chybu v dalších slovech a proto v rozpoznání jednotlivých slov může vznikat statistická závislost. Cílem této aktivity bude navrhnout metody pro porovnání výsledků rozpoznávání, které berou v úvahu statistickou závistost úspěšného rozpoznání slov ve větě.
Plánované indikátory dosažení - očekávané výsledky aktivity
Bude vyvinuta originální metodologie pro porovnání úspěšnosti výsledků rozpoznávání.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Dozažené výsledky budou prezentovány formou článků v odborném tisku a vystoupení na konferencích a seminářích.

Číslo aktivity
03/09
Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Trénování akustických a jazykových modelů

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Dosavadní výzkum byl zaměřen hlavně na vývoj a trénování akustických modelů a výběr nejlepších metod pro akustické modelování. Pro jazykové modelování byly použity gramatiky (pokud to bylo možné) nebo bylo provedeno testování úspěšnosti rozpoznávání bez použití jazykového modelu. Další výzkum se proto bude zabývat vývojem metod pro stochastické jazykové modelování. Součástí dalšího výzkumu bude rovněž testování použití akustických modelů založených na slabikách.

Plánované indikátory dosažení - očekávané výsledky aktivity

Natrénované akustické a jazykové modely.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Budou vyhodnoceny testy úspěšnosti rozpoznávání.

Číslo aktivity

04/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Vývoj pokročilého algoritmu pro automatickou sémantickou analýzu dotazů

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Cílem této aktivity je navázat na aktivitu 2008-03 (Vývoj základního algoritmu pro automatickou sémantickou analýzu dotazů) a vylepšit navržené algoritmy tak, aby bylo dosaženo vyšší přesnosti analýzy vět. Algoritmy budou modifikovány zejména zlepšováním použitého stochastického modelu a začleněním metod zpracování jazyka, které jsou specifické pro inflexní jazyky. Dále budou implementovány i jiné stochastické modely, které budou sloužit k porovnání výsledků dosažených algoritmem vyvinutým v rámci této aktivity.

Plánované indikátory dosažení - očekávané výsledky aktivity

Indikátorem dosažení výsledku je vyšší přesnost analýzy vět měřená procentem shody s testovací kolekcí vět.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky algoritmu budou publikovány v odborné literatuře. Programové nástroje budou uloženy v datovém úložišti projektu. Úspěšnost bude možno ověřit na vytvořených datech srovnáním výsledků algoritmu s výsledky lidských anotátorů.

Číslo aktivity

05/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Extrakce faktografických dat z veřejně dostupných zdrojů

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Navržený a otestovaný systém bude schopen na základě dotazu formulovaného v přirozené řeči poskytovat relevantní odpovědi. Nejprve bude nutné analyzovat strukturu typických dotazů formulovaných v přirozeném jazyce. Tato aktivita bude využívat data získaná v rámci aktivit 2007-37 (Pořizování korpusu dotazů z reálného prostředí) a 2008-01 (Sémantické anotování korpusu). Budou prozkoumány dva přístupy k řešení problému. V prvním budou využity výsledky aktivity 2008-03 (Vývoj základního algoritmu pro automatickou sémantickou analýzu dotazů). V druhém přístupu se použije systém generických šablon využívající lexikální analýzu. Výsledky přístupu budou statisticky srovnány.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem bude algoritmus extrakce dat a soubor generických šablon, měřitelným parametrem bude procentuální úspěšnost extrakce a porovnání obou přístupů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky budou prezentovány v odborné literatuře, software bude veřejně dostupný na internetových stránkách.

Číslo aktivity

06/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Ontologie, aplikace OWL a uživatelské rozhraní v oblasti ERP (Event related potentials)

Zahájení aktivity

1.10.2008

Ukončení aktivity

22.12.2009

Popis aktivity

Praktická realizace sémantického webu používá standardy a technologie RDF a OWL. Jejich úspěšné nasazení je však podmíněno důkladnou analýzou a modelováním vybrané domény. V rámci aktivity bude vytvořena ontologie pro doménu ERP (event-related potentials), vytvořen transformační mechanismus pro výměnu metadat s relační databází a finálně vytvořeno přívětivé uživatelské rozhraní.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem bude realizace ontologie ERP domény, transformační mechanismus pro výměnu metadat s relační databází a odpovídající uživatelské rozhraní.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky budou prezentovány v odborné literatuře, software bude veřejně dostupný na internetových stránkách.

Číslo aktivity

07/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Automatické rozpoznávání dialogových aktů na webových stránkách

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Tato aktivita navazuje na aktivity 2007-33 a 2008-15. Předchozí aktivity se zabývaly automatickým rozpoznáváním

dialogových aktů na řečovém korpusu. Cílem této nové aktivity je zaměřit se na rozpoznávání dialogových aktů z rozhovorů na webových stránkách. Je potřeba vyřešit následující problémy: 1) detekce částí webových stránek, kde se vyskytuje dialog 2) segmentace dialogu na dialogové jednotky (věty, dialogové akty) 3) automatické rozpoznání dialogových aktů. Na rozpoznání dialogových aktů budou použity zatím dostupné metody, které byly vytvořeny v rámci aktivit 2007-33 a 2008-15, avšak budou doplněny dalšími informacemi (např. větnou punktuací apod.). Vytvořené nástroje bude možno použít např. na automatickou anotaci korpusů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Funkční sada nástrojů, která bude umožňovat automatické rozpoznávání dialogových aktů z webových stránek.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ověření funkčnosti sady nástrojů při automatickém rozpoznávání dialogových aktů ve vybraných webových stránkách.

Číslo aktivity

08/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Kategorizace dokumentů neuronovou sítí

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

V počáteční fázi aktivity budou dokončeny rozsáhlé testy zaměřené na optimální nastavení parametrů sítě ART-2 používané pro kategorizaci dokumentů. V další fázi bude činnost zaměřena na návrh a implementaci neuronové sítě učené s učitelem, vhodné pro kategorizaci dokumentů. Navržená síť bude natrénována na dostatečném množství vstupních dokumentů a výsledky kategorizace budou porovnány s výsledky dosaženými v předchozích aktivitách, ve kterých byly pro kategorizaci použity sítě ART-2 a Kohonenova mapa. Vzhledem k tomu, že výsledky kategorizace jsou do značné míry ovlivněny transformací vstupního textu do číselné podoby (vytvoření tzv. kontextového vektoru), bude zvažována i možnost náhrady transformačního algoritmu algoritmem vhodnějším pro zpracování česky psaných dokumentů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem bude realizace neuronové sítě vhodné pro kategorizaci dokumentů, popř. modifikace algoritmu pro převod textové informace na číselný (tzv. kontextový) vektor vhodný pro další zpracování.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky budou prezentovány v odborné literatuře, software bude veřejně dostupný na internetových stránkách.

Číslo aktivity

09/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Zpracování korpusových záznamů (korpusy LAC-SS a LAC-Noise)

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Katalogizace záznamů v korpusech LAC-SS 2007 a 2008, jejich úpravy, předzpracování, transkripce a příprava k

využití pro trénování akustických modelů ASR systému JLASER, k přípravě databáze segmentů pro TTS systém jSynt a k návrhu a ověřování metod předzpracování akustického signálu pro potřeby ASR. Dále dokončení sběru dat do ruchového korpusu LAC-Noise, katalogizace a úpravy těchto záznamů a posléze jejich příprava k použití a využití při návrhu a ověřování metod zvýšení výkonu a spolehlivosti ASR systémů v obtížných příjmových podmínkách a také při zkoumání psychoakustických parametrů vnímání akustického signálu člověkem s cílem implementovat postupně získané poznatky do řetězce zpracování akustického signálu ASR systému (konkrétně JLASER).

Plánované indikátory dosažení - očekávané výsledky aktivity

Kvalitně připravený korpus pro použití k výše zmíněným aktivitám, opatřený transkripcí, příp. anotací, bezpečně uložený, dobře organizovaný a snadno systematicky přístupný. Kvalitní korpus je navíc i z ekonomického hlediska poměrně cenný materiál, takže ho lze i výhodně nabízet dalším výzkumným subjektům, ať už na komerční nebo výměnné bázi.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledky budou prezentovány v odborné literatuře, korpus bude veřejně dostupný na internetových stránkách pracoviště.

Číslo aktivity

10/09

Ke kterému dílčímu cíli se aktivita vztahuje

4 - Ověřování, testování a vyhodnocování testů navržených metod v reálném prostředí....

Název (cíl)aktivity

Ověření a testy modulů pro zpracování přirozeného zpracování jazyka v multilinguálním prostředí.

Zahájení aktivity

5.1.2009

Ukončení aktivity

30.6.2009

Popis aktivity

Aktivita bude pokračováním aktivity č. 10 z roku 2008. V rámci této aktivity budou testovány a ověřeny výsledky jednotlivých modulů pro zpracování přirozeného jazyka ve vícejazyčném prostředí. Srovnán a vyhodnocen bude vliv multilinguality textových dat na výsledky jednotlivých metod zpracování přirozeného jazyka. Důraz bude kladen na srovnání možností předzpracování textu do jazykově nezávislé formy.

Plánované indikátory dosažení - očekávané výsledky aktivity

Moduly verifikované standardními testy pro hodnocení kvality – přesnost, úplnost a statistická významnost výsledků metod.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Otestování nástroje na reálných datech na úlohách zpracování textu (např. klasifikace textu, disambiguace, sumarizace, vyhledávání, detekce plagiátů apod.)

Číslo aktivity

11/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Implementace systému pro vyhledávání a sumarizaci

Zahájení aktivity

5.1.2009

Ukončení aktivity

30.6.2009

Popis aktivity

Aktivita navazuje na výstupy z návrhu systému MUSE pro vyhledávání dokumentů v multilingválním prostředí. Prototypový systém bude optimalizován a provede se implementace funkčního modulu pro vyhledávání a sumarizaci.

Plánované indikátory dosažení - očekávané výsledky aktivity

Bude vytvořen funkční systém pro vyhledávání a sumarizaci ve vícejazyčném prostředí. Systém bude s dokumentací dodán jako autorizovaný software.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Otestování nástroje na reálných datech na úloze vyhledávání a sumarizace v multilingválním prostředí. Předání programu se provede v elektronické formě.

Číslo aktivity

12/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Metoda aktualizační sumarizace

Zahájení aktivity

5.1.2009

Ukončení aktivity

30.11.2009

Popis aktivity

V předchozí etapě projektu byla zkoumána sumarizace tématu. Cílem této aktivity je rozšířit sumarizační metodu o stanovení základních znalostí tématu. Souhrn by potom měl obsahovat pouze nové informace – aktualizační sumarizace (Update Summarization). Na vstupu budou dva shluky dokumentů pojednávajících o stejném tématu/události: dokumenty v prvním shluku uživatel již četl, dokumenty ve druhém ještě ne. Cílem je vytvořit souhrn „nových“ dokumentů, kde se nebudou vyskytovat informace uživateli již známé.

Plánované indikátory dosažení - očekávané výsledky aktivity

navržená metoda sumarizace bude implementována - vytvořen bude tzv. aktualizační sumarizátor.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Ověření kvality bude provedeno účastí na TAC (Text Analysis Conference - NIST). Výsledky vyvinutého sumarizátoru tak budou porovnány s výsledky ostatních skupin, které se zúčastní experimentů.

Číslo aktivity

13/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Online vyhledávací a sumarizační systém

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Cílem této aktivity bude rozšíření online sumarizačního systému SWEeT (http://tmrg.kiv.zcu.cz:8080/sweet) o možnost vytvářet aktualizační souhrny. Systém bude pracovat následovně: Uživatel vloží dotaz, který by měl být dostatečně bohatý, aby vymezil dané téma. Navíc zadá datum, které bude oddělovat redundantní informace (starší

dokumenty, s jejichž obsahem by měl být již seznámen) a nové informace (z dokumentů, jejichž datum je větší než zadaný). Vyhledané dokumenty pak zpracuje zabudovaný systém aktualizační sumarizace a výsledný souhrn, který by měl obsahovat pouze nové informace o daném tématu, bude vrácen uživateli.

Plánované indikátory dosažení - očekávané výsledky aktivity

Zpracování online systému veřejně přístupného z webu.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Systém bude testován několika uživateli. Bude zaznamenán jejich dotaz a odpověď systému. Anotována bude lexikální a obsahová kvalita souhrnu.

Číslo aktivity

14/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh a implementace pokročilé metody pro odhalování plagiátů s využitím LSA

Zahájení aktivity

1.7.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Navrhovaná aktivita volně navazuje na aktivitu č. 18 z roku 2008. V rámci této aktivity bude navržena a implementována pokročilejší metoda. Zájem bude věnován oblastem předzpracování textu a jeho vlivu na přesnost detekce plagiátů. Dále budou zahrnuty optimalizace výkonu pro rozsáhlejší data a otestovány nové matematické modely nad textovými dokumenty.

Plánované indikátory dosažení - očekávané výsledky aktivity

Implementace pokročilé metody využívající LSA pro odhalování plagiátů v psaném textu.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Navržená metoda bude odladěna a testována na relevanci výsledků s označkovaným českým korpusem z aktivity č. 16 z roku 2008. Výsledkem aktivity budou publikace na vědeckých konferencích.

Číslo aktivity

15/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Transformace zvolené ontologie do logického programu v jazyce Datalog

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Vytvoření ontologie pro datovou kolekci studijních programů, oborů a předmětů technických vysokých škol a univerzit v ČR, která byla vytvořena v rámci aktivity 2008-12. Transformace této ontologie a ontologie katastrof, která byla vytvořena v rámci téže aktivity, do formy logického programu založeného na pravidlově orientovaném jazyce Datalog. Oba programy budou využity k vyhodnocení typických dotazů prostřednictvím experimentálního deduktivního databázového systému. Stejné dotazy budou položeny nad příslušnou ontologií a výsledné odpovědi budou porovnávány.

Plánované indikátory dosažení - očekávané výsledky aktivity

Ontologie studijních programů a oborů a postup, jak tuto ontologii převést do logického jazyka Datalog. Očekává

se, že výsledný logický program může obsahovat i pravidla, která nebudou podchycena ontologií, ale budou vycházet ze závislostí uvedených v datové kolekci. Např. bude možno zjistit, zda (a jak) se liší skladba vybraného studijního programu na různých školách. Dalším výsledkem bude postup, jak převést ontologii katastrof do jiného logického programu, napsaného v jazyce Datalog. Použitá datová kolekce je tématicky zaměřená (na katastrofy), zde se neočekává výskyt dodatečných pravidel.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Výsledné transformované logické programy. Každý program by měl být ekvivalentní s odpovídající ontologií, tj. na položený dotaz poskytuje shodnou odpověď.

Číslo aktivity

16/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Zobecnění transformace ontologie do logického programu v jazyce Datalog

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Zobecnění transformačních pravidel převodu ontologie na logický pravidlově orientovaný program aplikovaných na ontologii z aktivity „Transformace ontologie do logického programu v jazyce Datalog“ (aktivita 2009-15). Stanovení vlastností výsledného logického programu, který vznikne transformací z ontologie.

Plánované indikátory dosažení - očekávané výsledky aktivity

Návrh a realizace obecných formalismů a algoritmů, které umožní transformaci ontologie na logický pravidlově orientovaný program. Stanovení vlastností, které splňuje výsledný logický program.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Navržený formalismus bude vytvářet takové logické programy, které budou ekvivalentní pro zadanou ontologií.

Číslo aktivity

17/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Rozšířené možnosti využití on-line slovníku SPOT

Zahájení aktivity

5.1.2009

Ukončení aktivity

22.12.2009

Popis aktivity

Cílem dílčího projektu SPOT pro následující období je rozšíření uživatelské základny a sběr zkušeností s takovým typem slovníku, plnohodnotná podpora pro překladatelské projekty a integrace s aplikacemi pro vyhledávání na sémantickém a sociálním webu. Detaily jsou dostupné na http://wiki.kiv.zcu.cz/SlovnikTerminologie/HomePage.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledek aktivity bude možno ověřit následujícími indikátory:

- provoz slovníku se zaregistrovanými a aktivními uživateli jak v roli přispěvatelů, tak editorů,
- množství překladatelských projektů využívajících slovník, a veřejně přístupná část korpusu vzniklá na základě jejich činnosti,
- počty dotazů do slovníku odeslané z jiných aplikací, s nimiž je integrován (vyhledávače, portály).

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Informace o počtu, aktivitě a projektech uživatelů budou dostupné z administrační části aplikace a budou moci být převedeny do formy statistik. Rozšířený korpus bude dostupný ve vyhledávací veřejné části aplikace spot.zcu.cz.

Číslo aktivity
18/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Metody automatického rozpoznávání dialogových aktů pro zpracování multimediálních vstupů
Zahájení aktivity
5.1.2009
Ukončení aktivity
22.12.2009
Popis aktivity
Tato aktivita navazuje na aktivity 2008-15 a 2008-22 a zabývá se automatickým rozpoznáváním dialogových aktů. Na rozdíl od dosavadního přístupu bude přistoupeno k integraci paralingvistických atributů dialogu, tzn. vstupem do systému rozpoznávání bude videosignál (zvuk i obraz). Takový model rozpoznávání dialogových aktů by měl značnou měrou přispět ke zvýšení jeho přesnosti. Cílem aktivity bude návrh nových metod automatického rozpoznávání dialogových aktů, které budou pracovat přesněji než metody existující a dále pak analýza účinnosti příznaků používaných pro rozpoznávání dialogových aktů. V tomto případě bude zvolen poněkud netradiční prostředek – datový sklad. Existující metody automatického rozpoznávání dialogových aktů využívají k určení aktuálního dialogového aktu kombinaci lexikálních (většinou v podobě posloupnosti slov ve větě) a prozodických příznaků. Lexikální metody modelují dialogové akty pomocí jazykových modelů typu n-gram. Naším cílem bude hlubší analýza prozodických příznaků a jejich vazba na výraz obličeje. Výraz obličeje nese velmi důležitou informaci o významu promluvy. Pro analýzu prozódie budou používány základní prozodické příznaky, v případě analýzy výrazu byla již navržena klasifikace do pěti základních tříd.
Plánované indikátory dosažení - očekávané výsledky aktivity
Výsledkem aktivity budou nové metody automatického rozpoznávání dialogových aktů využívající multimediálního vstupu a analýza účinnosti prozodických charakteristik. Metody budou prověřovány na existujících souborech dat získaných záznamem interaktivních dialogů v létech 2007-08, případně budou zaznamenány další alternativní formy vedení dialogů.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Navržené metody a výsledky analýzy budou diskutovány na pracovních setkáních, prezentovány na konferencích a publikovány ve vědeckých publikacích.

Číslo aktivity
19/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Úpravy a využití lexikální databáze VerbaLex obsahující valenční rámce českých sloves v algoritmech syntaktické a logické analýzy
Zahájení aktivity
1.1.2009
Ukončení aktivity
31.12.2009
Popis aktivity

Komplexní valenční rámce v databázi VerbaLex poskytují informace pro pokročilou syntaktickou a logickou analýzu české věty. V rámci dané aktivity budou navrženy a ověřovány konkrétní metody a algoritmy pro jejich využití v systému synt vyvíjenému v Centru ZPJ.

Plánované indikátory dosažení - očekávané výsledky aktivity

Nová verze lexikální databáze VerbaLex.

Implementované algoritmy v rámci syntaktické a logické analýzy, které využívají VerbaLexových rámců pro kvalitnější analýzu.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace

Číslo aktivity

20/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Návrh a vývoj nástrojů DEB platformy

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Nástroje platformy Dictionary Editor and Browser (DEB) mají široké využití, kde aktuálně vyvíjené systémy zahrnují aplikace pro Global WordNet Grid (vícejazyčná sémantická síť) a multiliguálně orientovanou lexikografickou stanici.

Plánované indikátory dosažení - očekávané výsledky aktivity

implementované rozšířené a nové nástroje a metody na platformě DEB

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace

Číslo aktivity

21/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Aplikace syntaktické analýzy pro určení syntaktických struktur ve velmi velkých korpusech

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Vyvíjený syntaktický analyzátor s využitím budovaného korpusu syntaktických stromů poskytuje na českých větách strukturní informace vyšší úrovně, které výrazně pomohou při inteligentní analýze rozsáhlých textů. V rámci dané aktivity bude systém syntaktické analýzy aplikován, testován a vylepšován vůči vytvářenému velmi velkému (stovky milionů pozic) českému korpusu.

Plánované indikátory dosažení - očekávané výsledky aktivity

Vylepšená analýza syntaktických struktur v českém textu na základě dat z velmi rozsáhlých korpusů.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

publikace

Číslo aktivity
22/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Rozšíření algoritmů logické analýzy českých vět
Zahájení aktivity
1.1.2009
Ukončení aktivity

Popis aktivity
Logická analýza věty umožní zpracovat základní sémantické znalosti a vztahy v dané větě. V rámci aktivity budou doplňována nová pravidla a navrženy a implementovány nové metody tvorby logické konstrukce české věty na základě její syntaktické analýzy, která využívá budované pokročilé jazykové zdroje jako jsou komplexní valenční rámce z databáze VerbaLex.
Plánované indikátory dosažení - očekávané výsledky aktivity
Vylepšené implementované algoritmy tvorby logické konstrukce pro dosud neanalyzovatelné gramatické fenomeny.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
publikace

Číslo aktivity
23/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Vytvoření velmi velkého českého korpusu
Zahájení aktivity
1.1.2009
Ukončení aktivity
31.12.2009
Popis aktivity
Pro ověřování úspěšnosti jednotlivých metod a algoritmů je vhodné mít co největší množství dat. Proto bude vytvořen velmi velký korpus obsahující české texty. Korpus bude anotován na různých úrovních, aby byl využitelný v co největším množství aplikací.
Plánované indikátory dosažení - očekávané výsledky aktivity
Korpus bude mít rozsah asi 1 miliardu tokenů, bude plně označkován na morfologické úrovni, budou v něm vyznačeny syntaktické vztahy jednotlivých slov.
Plánované prostředky ověření - forma zpracování a předání výsledku aktivity
Publikace na mezinárodních konferencích. Zpřístupnění korpusu uživatelům.

Číslo aktivity
24/09
Ke kterému dílčímu cíli se aktivita vztahuje
3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...
Název (cíl)aktivity
Systém pro analýzu anaforických vztahů
Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Aktivita si klade za všeobecný cíl implementovat program, který automaticky hledá a analyzuje anaforické vztahy ve volných textech.

Plánované indikátory dosažení - očekávané výsledky aktivity

Začlenění dalších zdrojů dat do procesu automatické analýzy
anaforických vztahů (AR). Analýza zvláštností anaforických vztahů v souvětích. Prozkoumání využitelnosti informací o aktuálním členění větném. Obecně: zlepšení úspešnosti automatické AR.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

článek na DAARC 2009

Číslo aktivity

25/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Klasifikace a podobnost matematických textů ve vytvořeném korpusu

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Aktivita se soustředí na využití shromážděného korpusu klasifikovaných matematických článků a analýzu jazyka matematiky v závislosti na tématice dokumentu. Budou provedeny experimenty s indexováním a vyhledáváním v matematickém korpusu a s~využitím podobnosti a klasifikace pro vyhledávání.

Plánované indikátory dosažení - očekávané výsledky aktivity

Prototyp aplikace navržené pro nasazení v České
digitální matematické knihovně DML-CZ.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

organizace workshopu a publikace v odborném sborníku

Číslo aktivity

26/09

Ke kterému dílčímu cíli se aktivita vztahuje

2 - Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka...

Název (cíl)aktivity

Rozšířená verze systému WebGen.

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Vylepšení použitých dialogových strategií na základě zpětné vazby od testerů. Návrh a implementace editace podporovaných stránek. Podpora pro další typy webových stránek. Propojení systému WebGen s anotovanou databází grafických objektů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem aktivity bude rozšířená verze systému WebGen s podporou editace stávajících prezentací a výstupy

testování systému WebGen a anotované databáze grafických objektů nevidomými uživateli.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Dosažená výsledky budou prezentovány na pracovních setkáních a formou publikací ve sbornících.

Číslo aktivity

27/09

Ke kterému dílčímu cíli se aktivita vztahuje

3 - Návrh a implementace metod komunikace s prostředím www v přirozeném jazyce. ...

Název (cíl)aktivity

Anotátor grafických objektů

Zahájení aktivity

1.1.2009

Ukončení aktivity

31.12.2009

Popis aktivity

Detailnější rozpracování architektury anotátoru, vytvoření grafických ontologií grafické databáze s příhlédnutím k potřebám systému. Bude dokončena první verze anotátoru grafických objektů.

Plánované indikátory dosažení - očekávané výsledky aktivity

Výsledkem aktivity bude anotovaná databáze grafických objektů. Dalším výsledkem bude prostředek sloužící k podpoře procesu anotace a dialogové rozhraní pro práci s databází. Propojení anotované databáze se systémem webgen.

Plánované prostředky ověření - forma zpracování a předání výsledku aktivity

Publikace ve sbornících, 1. verze anotátoru grafických objektů.

3.2.2. NÁVRH ZMĚN V ŘEŠENÍ PROJEKTU - rok 2009

Pč.	Typ	Popis
1	návrh změn v řešení projektu	Žádná změna v řešení projektu není plánována, avšak v průběhu řešení se opět mohou objevit další dílčí aktivity, které bude třeba vyřešit, aby mohlo být dosaženo plánovaných cílů projektu.

3.3. NÁKLADY PROJEKTU - rok 2009

3.3.1. NÁKLADOVÉ TABULKY ZA JEDNOTLIVÉ SUBJEKTY

Rok	2009
Typ	požadované
Organizace	Západočeská univerzita v Plzni
Role organizace	příjemce - koordinátor

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady požadované tis. Kč	z toho požadované z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**2965**	**2945**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**0**	**0**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**100**	**0**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**30**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**100**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**450**	**50**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**350**	**0**
F9. CELKEM	**3995**	**2995**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**3995**	**2995**	**0**	**1000**

Rok	2009
Typ	požadované
Organizace	Masarykova univerzita
Role organizace	spolupříjemce

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady požadované tis. Kč	z toho požadované z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**1850**	**1542**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**60**	**40**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**60**	**40**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**0**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**0**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**180**	**128**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**200**	**0**
F9. CELKEM	**2350**	**1750**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**2350**	**1750**	**0**	**600**

3.3.2. NÁKLADOVÁ TABULKA ZA PROJEKT

Rok	2009
Typ	požadované
PROJEKT	2C06009 - CELKEM

POLOŽKA UZNANÝCH NÁKLADŮ tis. Kč	Náklady požadované tis. Kč	z toho požadované z účelové podpory tis. Kč
F1. - Osobní náklady nebo výdaje na zaměstnance, kteří se podílejí na řešení projektu a jim odpovídající povinné zákonné odvody a případné příděly do FKSP	**4815**	**4487**
F2. - Náklady nebo výdaje na pořízení hmotného a nehmotného majetku (investice, kapitálové)	**0**	**0**
F3. - Náklady nebo výdaje na provoz a údržbu hmotného majetku používaného při řešení projektu	**60**	**40**
F4. - Další provozní náklady vzniklé v přímé souvislosti s řešením projektu	**160**	**40**
F5. - Náklady nebo výdaje na služby využívané v přímé souvislosti s řešením projektu	**30**	**0**
F6. - Náklady nebo výdaje na zveřejnění výsledků projektu včetně nákladů nebo výdajů na zajištění práv k výsledkům výzkumu	**100**	**0**
F7. - Cestovní náhrady vzniklé v přímé souvislosti s řešením projektu	**630**	**178**
F8. - Doplňkové (režijní) náklady nebo výdaje vzniklé v přímé souvislosti s řešením projektu, např. administrativní náklady, náklady na pomocný personál a infrastrukturu, enegii a služby neuvedené výše	**550**	**0**
F9. CELKEM	**6345**	**4745**

	PŘEVOD DO fondu tis. Kč	POUŽITÍ Z fondu tis. Kč
F0. - Zúčtování s Fondem účelově určených prostředků	**0**	**0**

	ZDROJE FINANCOVÁNÍ CELKEM tis. Kč	- z toho Účelová podpora (DOTACE) tis. Kč	- z toho Ostatní veřejné zdroje tis. Kč	- z toho Neveřejné zdroje tis. Kč
Z9.	**6345**	**4745**	**0**	**1600**

3.3.3. NÁVRH ZMĚN V NÁKLADECH - rok 2009

Pč.	Typ	Popis
*		

4. PŘÍLOHY

4.1. ZPRÁVA O POSTUPU ŘEŠENÍ PROJEKTU - rok 2008

4.1.1. POPIS ŘEŠENÍ PROJEKTU - seznam

	Pořadí	Soubor
	1	**Postup řešení projektu v roce 2008 - Plzeň** Soubor obsahuje přehled nejvýznamnějších výsledků řešení projektu dosažených v průběhu roku 2008. Všechny vytýčené cíle byly splněny, pro jejich naplnění však bylo třeba provést celou řadu dodatečných činností, které jsou z důvodu rozsáhlosti zprávy v odstavci popsány pouze částečně. Detailní výsledky je možno nalézt v přílohách. Zprava_2C06009_odst411.doc *(114 kB)*
	2	**Práce na projektu v rámci Centra ZPJ FI MU** Soubor obsahuje souhrnnou zprávu o aktivitách za r. 2008. zpravNPV208fin.rtf *(18 kB)*

4.1.2. DOSAŽENÉ VÝSLEDKY

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/01/2008**

Název výsledku

PageRank for Bibliographic Networks

Abstrakt

In this paper, we present several modifications of the classical PageRank formula adapted for bibliographic networks. Our versions of PageRank take into account not only the citation but also the co-authorship graph. We verify the viability of our algorithms by applying them to the data from the DBLP digital library and by comparing the resulting ranks of the winners of the ACM E. F. Codd Innovations Award. Rankings based on both the citation and co-authorship information turn out to be “better” than the standard PageRank ranking.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Byla vyvinuta nová metoda pro vyhodnocování autoritativnosti výzkumníků a výzkumných skupin, vycházející z metody PageRank. Metoda zohledňuje nejenom údaje získané pomocí klasické citační analýzy, ale modifikuje a zpřesňuje výsledky zahrnutím informací o spoluautorství. Výsledky objektivněji zobrazují publikační významnost hodnocených.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nová metoda se prokázala jako použitelná pro ranking výzkumu. Je obecně použitelná pro dolování ze struktury Webu. Dovoluje eliminovat při hodnocení vliv „citation lobby“.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ježek Karel Doc. Ing. CSc.**
Spojení	377632475 +420724236002 jezek_ka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
01	Fiala D., Rousselot F., Jezek K.: PageRank for Bibliographic Network. Scientometrics, vol.76, no. 1, pp. 135-158, 2008-12-18, ISSN 0138-9130, Akademiai Kiado, Springer	J - Článek v odborném periodiku	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/02/2008**

Název výsledku

Exploration and Evaluation of Citation Networks

Abstrakt

This paper enhance the PageRank formula modified and adapted for bibliographic networks. Our modifications of PageRank take into account not only the citations but also the co-authorship relationships. We verified the capabilities of the developed algorithms by applying them to the data from the DBLP digital library and subsequently by comparing the resulting ranks of the sixteen winners of the ACM SIGMOD E.F.Codd Innovations Award from the years 1992 till 2007. Such ranking, which is based on both the citation and co-authorship information, gives better and more fair-minded results than the standard PageRank gives. The proposed method is able to reduce the influence of citation loops. The possibilities for farther improvements are suggested e.g. introducing temporal views into the citations evaluating algorithms.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Byla reformulována metoda sdružující metodu citační analýzy se sítí spoluautorství a rozšířena zkušební kolekce. Byly navrženy možné směry výzkumu vedoucí k další objektivizaci výsledných hodnocení.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Byla zformulována a zdůvodněna metoda pro ranking výzkumu, včetně návrhu jejího dalšího rozvíjení.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ježek Karel Doc. Ing. CSc.**
Spojení	377 632 447 +420724236002 jezek_ka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
02	Jezek K., Fiala D., Steinberger J.: Exploration and Evaluation of Citation Network. In Proceedings of the 12th International Conference on Electronic Publishing, ISBN 978-0-7727-6315-0, pp 351-362, Toronto, Canada 2008	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/03/2008**

Název výsledku

Automatic Text Summarization (The state of the art 2007 and new challenges)

Abstrakt

The headline of this paper names a research area originating from the late 50's but not loosing its popularity until the present time. Moreover, one of the most relevant today's problems caused by the rapid growth of the Web, which is called information overloading, has increased the necessity of more sophisticated and powerful summarizers. This paper shortly introduces a taxonomy of summarization methods and an overview of their principles from classical ones, over corpus based, to knowledge rich approaches. We consider various aspects which can affect their classification. A special attention is devoted to application of recent information reduction methods, based on algebraic transformations. Further, we introduce experiences with the development of our own summarizing method. Finally, some new ideas and a conception for the future of this field are mentioned.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Ohodnocení vlastností sumarizačních metod se zvláštním důrazem na algebraické metody pro redukci informací. Prezentace zkušeností s realizovaným sumarizátorem. Dále jsou diskutovány koncepce a možnosti dalšího rozvoje tohoto perspektivního oboru.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Odborná komunita byla seznámena s aktuálním stavem výzkumu a s reálnými možnostmi v oblasti sumarizace textů. Kriticky byly posouzeny směry dalšího zkoumání a vytýčeny cíle a metody následujícího postupu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ježek Karel Doc. Ing. CSc.**
Spojení	377632475 +420724236002 jezek_ka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
03	Ježek K., Steinberger J.: Automatic Text Summarization (The state of the art 2007 and new Challenges). In Proceedings of Znalosti 2008, Bratislava, Slovakia, February 2008, pp. 1–12, ISBN 978-80-227-2827-0.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/04/2008**

Název výsledku

SPOT - Slovník překladů odborné terminologie

Abstrakt

Slovník překladů odborné terminologie. Cílem projektu slovníku SPOT je pomoci překladatelům a všem zájemcům v úsilí o vytváření a používání korektních překladů složitých a/nebo nových termínů. Vytvořená aplikace podporuje tyto činnosti v podobě webového rozhraní, které obsahuje veřejnou a administrační část a zajišťuje oddělení práv uživatelů (zejména registrovaný uživatel vs. editor korpusu) pomocí rolí.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Všechny slovníkové aplikace dostupné na webu se soustředí na vyhledávání v pevném korpusu, případně umožňují jeho rozšiřování editorskou radou. SPOT přidává možnost zacílení na specifickou doménu, zapojení odborné komunity a integraci diskusních mechanismů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Kompletní software pro provoz a správu slovníku, který je možno použít pro libovolný slovníkový korpus včetně jeho dalšího rozšiřování.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Brada Přemysl Ing PhD. MSc.**
Spojení	377632435 brada@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň spot.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
04	Dokumentace je dostupná na http://www.kiv.zcu.cz/vyzkum/software/	R – software (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/05/2008**

Název výsledku

Plagiarism Detection based on Singular Value Decomposition

Abstrakt

Plagiarism is a widely spread problem that is the main focus of interest these days. In this paper, we propose a new method solving associations of phrases contained in text documents. This method, called SVDPlag, employs Singular Value Decomposition (SVD) for this purpose. Further, we discuss other approaches to plagiarism detection and compare them with our method. To examine the efficiency of plagiarism detection methods, we used an experimental corpus of 950 text documents about politics, which were created from the standard CTK corpus. The experiments indicate that our approach significantly improves the accuracy of plagiarism detection and overcomes other methods.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Návrh nové metody pro automatickou detekci plagiátů s využitím singulární dekompozice

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Metoda s využitím singulární dekompozice dosahuje vyšší přesnosti detekce plagiátů než ostatní běžně používané metody.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
05	Ceska, Z.: „Plagiarism Detection based on Singular Value Decomposition“. Advances in Natural Language Processing, LNCS/LNAI 5221, pp. 108-119, Springer Verlag Berlin Heidelberg, the 6th International Conference on Natural Language Processing (GoTAL 2008), Gothenburg, Sweden, August 2008. ISSN 0302-9743. ISBN 978-3-540-85286-5.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/06/2008**

Název výsledku

Multilingual Plagiarism Detection

Abstrakt

Multilingual text processing has been gaining more and more attention in recent years. This trend has been accentuated by the global integration of European states and the vanishing cultural and social boundaries. Multilingual text processing has become an important field bringing a lot of new and interesting problems. This paper describes a novel approach to multilingual plagiarism detection. We propose a new method called MLPlag for plagiarism detection in multilingual environment. This method is based on analysis of word positions. It utilizes the EuroWordNet thesaurus which transforms words into language independent form. This allows to identify documents plagiarized from sources written in other languages. Special techniques, such as semantic-based word normalization, were incorporated to refine our method. It identifies the replacement of synonyms used by plagiarists to hide the document match. We performed and evaluated our experiments on monolingual and multilingual corpora and results are presented in this paper.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Návrh metody pro automatickou detekci plagiátů ve vícejazyčném prostředí. Konkrétně bylo otestováno na jazycích čeština/angličtina.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Možnost odhalit překlady z anglických zdrojů. Příkladem může být například časté kopírování studentských prací z anglické Wikipedie a překlad do českého jazyka.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing.**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
06	Ceska, Z., Toman, M., Jezek, K.: „Multilingual Plagiarism Detection“. Artificial Intelligence: Methodology, Systems, and Applications, LNCS/LNAI 5253, pp. 83-92, Springer-Verlag Berlin Heidelberg, the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2008), Varna, Bulgaria, September 2008. ISSN 0302-9743. ISBN 978-3-540-85775-4.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/07/2008**

Název výsledku

Extrakce N-gramů z rozsáhlých textů

Abstrakt

V úlohách zpracování přirozeného jazyka jsou k reprezentaci textových dokumentů nejčastěji používána jednotlivá slova. Celkové výsledky lze však často vylepšit použitím dalších, sofistikovanějších položek. Mezi ně patří i N-gramy, pro jejichž extrakci byly publikovány algoritmy založené na různých principech. Existující techniky však nejsou primárně určeny pro zpracování velkého objemu dat, což je v současné době zásadní požadavek. V tomto článku prezentujeme algoritmus pro extrakci N-gramů z rozsáhlých textových korpusů. Srovnání s jinými přístupy naznačují, že naše řešení dosahuje výrazně lepších výsledků s ohledem na čas a množství zpracovaných dat.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Optimalizovaná metoda umožňující extrakci N-gramů z rozsáhlých textových dat na jednom počítači bez použití specializovaného hardware. Námi navržená metoda překonává ostatní metody ve smyslu časových i paměťových požadavků a splňuje kritéria kladená na množství zpracovaných dat na webu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přínosy jsou ekonomického charakteru. Námi navržená metoda umožňuje rychlé zpracování rozsáhlých dat na jednom počítači, čímž podstatně snižuje náklady na hardwarové vybavení.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing.**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
07	Ceska, Z., Hanak, I., Tesar, R.: „Extrakce N-gramů z rozsáhlých textů". Proceedings of the 7th Annual Conference ZNALOSTI 2008, Bratislava, Slovakia, pp. 54-65, February 2008. ISBN 978-80-227-2827-0.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/08/2008**

Název výsledku

Využití moderních přístupů pro detekci plagiátů

Abstrakt

Plagiátorství je v současnosti nejvíce skloňovaným pojmem, se kterým se můžeme setkat v každé oblasti lidské tvůrčí práce. Školství je jednou z důležitých oblastí, kde je nutné tomuto zamezit. V tomto článku se zabýváme moderními přístupy pro detekci plagiátů textových dokumentů. Naše metoda využívá normalizaci textu a latentní sémantickou analýzu pro nalezení skrytých vztahů mezi dokumenty. Dále uvádíme předběžné experimenty provedené na testovacím korpusu, který obsahuje 950 textových dokumentů o politice. Předběžné experimenty naznačují výhodnost naší metody a zlepšení výsledků oproti ostatním přístupům. V závěru článku diskutujeme využití WordNet tezauru pro zlepšení přesnosti současných metod a možnosti identifikace plagiátů, které byly přeloženy do jiných jazyků.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Metoda pro automatickou detekci plagiátů s využitím singulární dekompozice. Diskuse možností pro využití WordNet tezauru a identifikace překladu do jiných jazyků.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Využití WordNet tezauru v oblasti automatické detekce plagiátů a aplikace na metodu singulární dekompozice vztahů frází.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing.**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
08	Češka, Z.: „Využití moderních přístupů pro detekci plagiátů“. Proceedings of the ITAT 2008, Information Technologies – Applications and Theory, Hrebienok, Slovakia, pp. 23-26, September 2008. ISBN 978-80-969184-8-5.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/09/2008**

Název výsledku

Free-Text Plagiarism Detection Based on Latent Semantic Analysis

Abstrakt

Plagiarism is a widely spread problem that is the main focus of interest these days. In this work, I describe the state of the art of free-text plagiarism detection methods. Further, I discuss approaches for text pre-processing and N-gram extraction that essentially influence the effectiveness of copy detection methods. I propose an advanced plagiarism detection method based on Latent Semantic Analysis (LSA). This method can employ two different document model representations and four variants for model factorization, such as Singular Value Decomposition (SVD), High Order Singular Value Decomposition (HOSVD), Non negative Matrix Factorization (NMF), and Non negative Tensor Factorization (NTF). The final goal is to propose a new method to be more effective than others do.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Popis existujících řešení pro automatickou detekci plagiátů, jejich výhody a nevýhody. Návrh nové metody využívající latentní sémantickou analýzu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Diskuse současných a nových řešení pro automatickou detekci plagiátů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Češka Zdeněk Ing.**
Spojení	377 632 452 zceska@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
09	Ceska, Z.: „Free-Text Plagiarism Detection Based on Latent Semantic Analysis“. Technical Report No. DCSE/TR-2008-01, Pilsen, Czech Republic, April 2008.	O - Ostatní výsledky, které nelze zařadit do žádného z výše uvedených druhů výsledku	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/10/2008**

Název výsledku

Srovnání přístupů extrakce užitečné informace z webu

Abstrakt

V článku srovnáváme dvě metody pro extrakci užitečné informace z webu. První metoda je založena na statistické analýze struktury webové stránky a druhá metoda využívá dotazy XQuery pro extrakci informace z částečně strukturovaných dokumentů. V testech srovnáváme přesnost a úplnost automatické extrakce pomocí obou metod a ručně vytvářeného referenčního extraktu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Metoda XQT produkuje přesné výsledky s možností jemného strukturování výsledných dat. Je vhodná pro tvorbu textových korpusů a extrakci obecných dat s důrazem na přesnost. Metoda NIT poskytuje výsledky s přesností a úplností pohybující se kolem 80 %. Použití metody je jednoduché, protože kvalita extrakce závisí na jediném parametru. Metoda je určená výhradně pro tvorbu textových korpusů z webových zdrojů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Vytvořili jsme dvě metody pro extrakci textu z webových stránek a provedli jejich testy na datových korpusech.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Toman Michal Ing.**
Spojení	377 632 452 mtoman@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
10a	Toman, M.: Srovnání přístupů extrakce užitečné informace z webu, Proceedings of the 7th Annual Conference ZNALOSTI 2008, Bratislava, Slovakia, February 2008, 389-392. ISBN 978-80-227-2827-0.	D – článek ve sborníku (RIV 2009)	CES
10b	TOMAN, M.: Comparison of Approaches for Information Extraction from the Web, Proceedings of the 9th international PhD Workshop on Systems and Control . Ljubljana : Jožef Stefan Institute, 2008. s. 1-3. ISBN 978-961-264-003-3.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/11/2008**

Název výsledku

Web Topic Summarization

Abstrakt

In this paper, we present our online summarization system of web topics. The user defines the topic by a set of keywords. Then the system searches the Web for the relevant documents. The top ranked documents are returned and passed on to the summarization component. The summarizer produces a summary which is finally shown to the user. The proposed architecture is fully modular. This enables us to quickly substitute a new version of any module and thus the quality of the system's output will get better with module improvements. The crucial module which extracts the most important sentences from the documents is based on the latent semantic analysis. Its main property is independency of the language of the source documents. In the system interface, one can choose to search a news site in English or Czech. The results show a very good search quality. Most of the retrieved documents are fully relevant, only a few being marginally relevant. The summarizer is comparable to state-of-the-art systems.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Popis architektury systému SWEeT – online vyhledávání dokumentů relevantních k dotazu a jejich sumarizace

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Propojení vyhledávání textů a jejich sumarizace

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Steinberger Josef Ing. PhD.**
Spojení	377632431 jstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň textmining.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
11	Josef Steinberger, Karel Ježek, Martin Sloup: Web Topic Summarization, Proceedings of the 12th International Conference on Electronic Publishing, pp 322-334, Toronto, Canada 2008, ISBN 978-0-7727-6315-0	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/12/2008**

Název výsledku

Stochastic Parsing in a Hybrid Semantic Analysis System

Abstrakt

This article is focused on the problem of meaning recognition in spoken utterances. The goal is to find a computer algorithm capable to construct the meaning deion of a given sentence. Such an algorithm is used in applications that require human computer interaction in a natural language. In this article we describe some experiments that we made in this area of computation linguistic research.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

New approaches to the semantic analysis were described in the article.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The proposed methods allow robust semantic analysis of user utterances with high accuracy.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing.**
Spojení	377 632 491 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
12	Konopík, M. Habernal, I. - Stochastic Parsing in a Hybrid Semantic Analysis System, In Proceedings of 9th International PhD workshop on Systems and Control, Slovenia, 2008. ISBN 978-961-264-003-3	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/13/2008**

Název výsledku

LINGVOParser

Abstrakt

Vyvinutá metoda sémantické analýzy je založena na bezkontextových gramatikách obsahujících sémantické tagy. Vyvinuli jsme jednoduchý a účinný mechanismus na propagování informace z tagů za použití instrukcí, tzv. aktivních tagů.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Vytvořený software používá nově navržený formalismus aktivních tagů pro snadné získání sémantického obsahu z parsovacího stromu. Knihovna umožňuje díky aktivním tagům jednoduché použití sémantických gramatik pro složitěji strukturované vstupní věty. Efektivita vývoje gramatik je vyšší díky nezávislosti na určitém skriptovacím jazyce.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Knihovna slouží k sémantické analýze vět. Je snadno použitelná v projektech pro zpracování přirozeného jazyka.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Habernal Ivan Ing.**
Spojení	377 632 491 habernal@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
13a	Habernal, I. Konopík, M. - Active Tags for Semantic Analysis, In Text, speech and dialogue 2008. Berlin : Springer, 2008. s. 69-76. ISSN 0302-9743. ISBN 978-3-540-87390-7	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG
13b	Software ke stažení http://www.kiv.zcu.cz/vyzkum/software/detail.html?id=22	S - Prototyp, uplatněná metodika, funkční vzorek, autorizovaný software, výsledky aplikovaného výzkumu promítnuté do právních předpisů a norem, užitný vzor	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/14/2008**

Název výsledku

Active Tags for Semantic Analysis

Abstrakt

A new method for semantic analysis is proposed in this paper. The method is based on handwritten context-free grammars enriched with semantic tags. We developed an easy-to-use and yet very powerful mechanism for tag propagation. The mechanism allows the semantic information to be easily extracted from the parse tree. The propagation mechanism is based on an idea to add propagation instruction to the semantic tags. The tags with such instructions are called active tags in this article. Using the proposed method we developed a useful tool for semantic parsing that we offer for free on our internet pages.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Associating the rules of a context-free grammar with semantic tags is beneficial however, after parsing the tags are spread across the parse tree and it is usually hard to extract the complete semantic information from it. The mechanism of tag merging and propagation of semantic information using the active tags allows to keep the grammar independent on the target programming or ing language.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

We found in our experiments that this approach is extremely helpful in building semantic analysis applications. The proposed method is used as a semantic parsing algorithm in a voice-driven chess game. We also develop a semantic analysis algorithm for spoken queries to an internet search engine which uses the active tags formalism for so-called local parsing.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Konopík Miloslav Ing**
Spojení	377 632 491 konopik@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
14	Habernal, I. Konopík, M. - Active Tags for Semantic Analysis, In Text, speech and dialogue 2008. Berlin : Springer, 2008. s. 69-76. ISSN 0302-9743. ISBN 978-3-540-87390-7.	D - Článek ve sborníku z akce (publikovaná přednáška – proceeding)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/15/2008**

Název výsledku

Lexical Class Semantic Analysis

Abstrakt

Semantic analysis of lexical classes is a fundamental step of semantic analysis based on stochastic semantic parsing. The lexical class is a single word or a word group with specific semantic information such as dates, times, cities, etc. Having obtained a set of lexical classes, a semantic parse tree can be built upon it. This tree describes the relation between lexical classes and their appropriate superior concepts. This paper describes an implementation of a lexical class identification based on context-free grammars and parsing methods. The semantic analysis of lexical classes is based on grammars enriched with semantic tags. The main algorithm is described together with the experimental results.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

The "proof of concept" for lexical class identification and semantic analysis was developed and tested. Parsing methods based on active bottom-up chart parser and context-free grammars seem to be an efficient approach to the lexical class identification. The context-free grammars are able to cover more complicated lexical classes such date and time.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The fundamental step of a hybrid stochastic semantic parsing was developed. It yields very good results on testing data.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Habernal Ivan Ing**
Spojení	377 632 491 habernal@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
15	Habernal, I. Konopík, M. - Lexical Class Semantic Analysis, In Proceedings of 9th International PhD workshop on Systems and Control, Slovenia, 2008. ISBN 978-961-264-003-3	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/16/2008**

Název výsledku

Categorization of Czech written documents using WEBSOM methods

Abstrakt

The method called WEBSOM was designed for automatic processing and categorization of English and Finnish written documents and the following information retrieval in these documents. We applied this method (based on two layer architecture) to categorization of Czech written documents. Our research was focused on the syntactic and semantic relationship within word categories of word category map (WCM) and on the results provided by document category map (DCM) with respect to the content of WCM. The document classification system was tested on a subset of 100 documents (manual work was necessary) from the corpus of Czech News Agency documents. The result confirmed that not only WEBSOM method but also humans have problems with natural language semantics and determination of semantic domains from word categories.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

The behavior of WEBSOM method was tested on the set of Czech documents. Syntactic and semantic relationships within word categories were investigated in detail. There were significant differences between students undergoing the semantic experiment.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The results obtained by application of WEBSOM method to a collection of Czech written documents confirmed a general problem connected with document semantics and document classification. Not only WEBSOM method but also humans have problems with classification of word categories into semantic domains.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Mouček Roman Ing. PhD.**
Spojení	377 632 465 moucek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
16	Mouček, R., Mautner, P.: Categorization of Czech written documents using WEBSOM methods, In: Proceedings of 9th International PhD workshop on Systems and Control, Slovenia, 2008. ISBN 978-961-264-003-3.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/17/2008**

Název výsledku

Zpracování a kategorizace česky psaných dokumentů neuronovou sítí

Abstrakt

Tento článek se zabývá aplikací metody WEBSOM na kolekci česky psaných dokumentů. Je zde popsán základní princip metody, způsob převodu textové informace na číselnou reprezentaci zpracovávanou Kohonenovou mapou. Alternativně s Kohonenovou mapou byla testována i Carpenter-Grossbergova ART-2 síť, běžně používaná pro adaptivní shlukování vstupních vektorů. Výsledky dosažené s využitím této sítě jsou rovněž prezentovány v tomto článku.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Nový přístup ke kategorizaci česky psaných dokumentů

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Testované metody umožnují využití neuronových sítí pro zpracování a kategorizaci česky psaných dokumentů. Implementovaný software je snadno použitelný pro zpracování různých textových korpusů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Mautner Pavel Ing. PhD.**

Spojení mautner@kiv.zcu.cz

Organizace 49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
17	MAUTNER, P. MOUČEK, R. Zpracování a kategorizace česky psaných dokumentů neuronovou sítí. In Informatika v škole a v praxi . Ružomberok : Pedagogická fakulta Katolíckej univerzity, Slovensko, 2008, ISBN 978-80-8084-362-5.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/18/2008**

Název výsledku

Sémantika přirozeného jazyka a reálného světa – počítačové zpracování.

Abstrakt

Článek se zabývá možnostmi počítačového zpracování sémantiky přirozeného jazyka a reálného světa a pokládá otázku, do jaké míry je toto zpracování možné a smysluplné. Odpověď pak hledá v kombinaci poznatků a zkušeností tří různých oborů - neurověd, lingvistiky a informatiky. Stručně je prezentován pohled neurověd na fungování lidského mozku a jsou popsány paměťové složky mající vliv na zpracování sémantiky přirozeného jazyka a sémantiky vnějšího reálného světa. Krátce je představen i vnější lingvistický pohled na přirozený jazyk a jeho sémantické roviny. Z informatických oborů jsou pak představeny přístupy umělé inteligence a softwarového inženýrství. Zmíněna je i vize sémantického webu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Sémantika je zkoumána z pohledu tří různých oborů - neurověd, lingvistiky a informatiky kombinace těchto poznatků pak vytváří nový pohled na problematiku zpracování sémantiky přirozeného jazyka

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Bližší stanovení mezí a možností při zpracování sémantiky přirozeného jazyka. Lepší odhad smysluplnosti vývoje automatických metod počítačového zpracování jazyka.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Mouček Roman Ing. PhD.**
Spojení	377 632 465 moucek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
18	Mouček, R., Mautner, P.: Sémantika přirozeného jazyka a reálného světa – počítačové zpracování, Sborník 4. ročníku mezinárodní konference „Informatika v škola a v praxi“, Ružomberok, Slovensko, 2008, ISBN 978-80-8084-362-5.	D – článek ve sborníku (RIV 2009)	CES

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/19/2008**

Název výsledku

Akustický model založený na neuronové síti a decision tree clusteringu.

Abstrakt

Článek se pokouší porovnat výkon neuronových sítí a směsí Gausových funkcí při použití pro automatické rozpoznávání řeči. V naší práci jsme použili vícevrstvý perceptron pro odhad emisních pravděpodobností skrytého Markovova modelu. Stejně jako v případě směsí Gaussových funkcí je i u neuronových sítí třeba řešit problém nedostatku trénovacích dat pro málo frekventované fonetické jednotky. Toho je docíleno tzv. svazováním stavů, přičemž rozhodnutí, které stavy svázat zajišťuje shlukovací algoritmus založený na rozhodovacích stromech. Celkový počet svázaných stavů je možné ovlivnit nastavením prahu shlukovacího algoritmu. Ukazuje se, že využití neuronových sítí může vést k lepším výsledkům, pokud je výsledný počet stavů dostatečně malý.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Jedná se o nový způsob, jak využít neuronové sítě pro modelování kontextově závislých fonetických jednotek.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Využití kontextově závislých fonetických jednotek výrazně zlepšuje úspěšnost rozpoznávání.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing.**
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
19	Pavelka, T., Král, P.: Neural Network Acoustic Model with Decision Tree Clustered Triphones, Proceedings of 2008 IEEE International Workshop on Machine Learning for Signal Processing, ISBN 978-1-4244-2376 , Cancun, Mexico, 2008	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/20/2008**

Název výsledku

JLASER - Automatický rozpoznávač řeči

Abstrakt

Implementace automatického rozpoznávače řeči v jazyce Java. Podporuje MFCC parametrizaci a akustické modely založené na neuronových sítích nebo směsích Gaussových funkcí. Rozpoznávač je částečně kompatibilní s hojně používaným systémem pro trénování skrytých Markovových modelů CUED HTK (umožňuje načítat modely založené na směsích Gaussových funkcí a slovní sítě vytvořené z gramatik). Jako hlavní zdroj lexikální a syntaktické informace slouží orientovaný graf, který reprezentuje skrytý Markovův model. JLASER poskytuje funkce pro výrobu těchto grafů a to včetně tzv. lexikálních stromů, které výrazně zrychlují rozpoznávání. Dále je k dispozici efektivní dekodér, který hledá nejpravděpodobnější cestu tímto grafem (z té je možné extrahovat slova promluvy).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Mimo tradiční metody založené na skrytých Markovových modelech a směsích Gaussových funkcí podporuje tzv. hybridní přístup založený na umělých neuronových sítích.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Implementace v jazyce Java zajišťuje snadnou přenositelnost a návaznost na další projekty skupiny LIKS.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing.**
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
20	Software ke stažení: http://www.kiv.zcu.cz/vyzkum/software	R – software (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/21/2008**

Název výsledku

LASER Recorder (LRecorder)

Abstrakt

LRecorder je software pro automatizovaný, scénářem řízený sběr korpusového materiálu do zvukových korpusů, užívaných pro trénování ASR systémů. Umožňuje automatizovaný sběr záznamů do zvukových korpusů, zjednodušuje nahrávání podle vytvořeného scénáře a automaticky ukládá a katalogizuje nahrávky tak, že je lze okamžitě použít pro trénování ASR systému.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Zvýšení efektivity práce při sběru korpusového materiálu, neboť odpadá značná část činností potřebných při sběru materiálu pomocí běžného záznamového softwaru.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Zvýšení efektivity práce při sběru korpusového materiálu, neboť odpadá značná část činností potřebných při sběru materiálu pomocí běžného záznamového softwaru.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Ekštein Kamil Ing. PhD.**
Spojení	377 632 491 kekstein@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
21	Software ke stažení: http://www.kiv.zcu.cz/vyzkum/software	R – software (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/22/2008**

Název výsledku

N-Best Decoder for the JLASER Automatic Speech Recognizer

Abstrakt

The most common method for automatic speech recognition are the hidden Markov models (HMMs) where the most likely word sequence is found by the Viterbi algorithm. One of the main problems of this approach is, that knowledge sources that violate the first order Markov assumption cannot be used during the search. One solution to this problem is the so called N-Best paradigm which modifies the Viterbi algorithm so that it can return a list of ordered highest scoring state sequences. These can later be reordered by a search with non Markovian knowledge sources such as e.g. semantics. This article discusses the design and testing of the N-best decoding algorithms used in the JLASER recognizer. We have implemented two algorithms: one that guarantees to find the N-best solution and one faster but approximate algorithm. Our results show that the use of the sub-optimal algorithm did not lead to degradation in recognition accuracy while it greatly increased speed.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

It can be seen from the experimental results that the use of an approximate method (Lattice N-Best) did lead to increase in speed but did not degrade recognition accuracy.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

The new decoding algorithm allows the system to return more than one answer which can be beneficial in later stages of speech processing.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing.**
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
22	Pavelka, T., Brychcín, T.: N-Best Decoder for the JLASER Automatic Speech Recognizer, Proceedings of 9th International PhD Workshop on Systems and Control (YGV2008), ISBN: 978-961-264-003-3, Izola, Slovenia, 2008.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/23/2008**

Název výsledku

Automatic speech recognition using context-dependent syllables

Abstrakt

In this work, we deal with advanced context-dependent automatic speech recognition (ASR) of Czech spontaneous talk using hidden Markov models (HMM). As we have shown in previous works context-dependent units (e.g. triphones, diphones) in ASR systems provide significant improvement against simple non-context-dependent units. However, the usage of triphones brings some problems that we must solve. Mainly it is the total number of such units in the recognition process. To overcome problems with triphones we experiment with syllables. Syllables in the Czech spoken language are the smallest units recognisable by human. More over syllables are context-dependent units and their number is much lower than the number of triphones. Using our syllabification process we generate a list of units for recognition from our trainging corpus. Thanks to a slightly modified decision-tree clustering (used in previous works) we can even recognize units not included in the training corpus. The main part of this article shows problems with the implementation of syllables into the LASER (ASR system developed at Department of Computer Science and Engineering, Faculty of Applied Sciences) and results of the recognition process. These results are very promising. According to preliminary results we see better time performance and better accuracy of this new recognizer based on syllables.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

We have successfully built and tested a unique acoustic model set based on syllables.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

We have proved that syllables can be used as acoustic models. This leads to advantages in recognizer and opens new oportunities in models clustering.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Hejtmánek Jan Ing.**
Spojení	377 632 491 hejtman2@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
23	Hejtmánek, J., Pavelka, T.: Automatic Speech Recognition Using Context-dependent Syllables, Proceedings of 9th International PhD Workshop on Systems and Control (YGV2008), ISBN 978-961-264-003-3, Izola, Slovenia, 2008.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/24/2008**

Název výsledku

Hybrid Methods of Automatic Speech Recognition

Abstrakt

The thesis explores the possibilities of acoustic modeling by neural networks in automatic speech recognition. Today"s most used approach to speech recognition is the application of the theory of hidden Markov models (HMMs). When used for recognition of speech the emission probabilities are commonly modeled by mixtures of Gaussian functions. Research suggests that this task can also be done by neural networks. This is often referred to as the hybrid approach and experiments have shown that it can be advantageous in comparison with Gaussian mixture models (GMMs). The first phase of our work was to take each type of acoustic model and try to achieve the highest possible recognition accuracy. Experiments were carried out to find how the resulting accuracy changes with the number of trainable parameters (i.e. the number of Gaussians per mixture or the number of neurons in the network) and what changes can be made in the training process. Our previous research into Gaussian mixture models shows that a significant performance increase can be made by introducing context dependent phonetic units, namely decision tree clustered triphones. In this work we have attempted to do the same with neural networks. The next part of the work was to try to compare both kinds of acoustic models in terms of recognition accuracy and speed. We have found that in order to do so the models have to be tested as an integral part of a recognition system. Tests done with context independent phonetic units have demonstrated that neural networks use their trainable parameters more efficiently than their GMM counterparts. This leads to higher recognition speeds. However, significant increase in recognition accuracy can be achieved by utilizing context dependent phonetic units where neural networks have been applied only with limited success: We have found that neural networks can perform better than GMMs only if the total number of clustered triphones is kept low.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Využití neuronových sítí pro automatické rozpoznávání řeči, návrh metodologie pro porovnání různých typů akustických modelů, návrh efektivního dekodéru.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

V rámci této práce byla natrénována většina akustických modelů používaných rozpoznávačem JLASER.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pavelka Tomáš Ing.**
Spojení	377 632 491 tpavelka@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
24	Pavelka, T.: Hybrid Methods of Automatic Speech Recognition, PhD. Thesis, University of West Bohemia, Pilsen, 2008	V – výzkumná zpráva (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/25/2008**

Název výsledku

Evaluation of Dialogue Act Recognition Approaches

Abstrakt

This paper deals with automatic dialogue act recognition. Dialogue acts (DAs) are utterance-level labels that represents different states of a dialog, such as questions, statements, hesitations, etc. Information about actual DA can be seen as the first level of dialogue understanding. The main goal of this paper is to compare our dialogue act recognition approaches that model the utterance structure, and are particularly useful when the DA corpus is small, with n-gram based approaches. Results of our best approach are also combined with prosody. We show that our approaches significantly outperform the n-gram based methods. When prosody is used, the recognition accuracy is also slightly increased.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Testování nových metod pro automatické rozpoznávání dialogových aktů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Prakticky jsme ověřili, že námi navržené metody, které modelují globální větnou strukturu, pracují s výrazně vyšší přesností, než metody založené na n-gramech.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing. PhD.**
Spojení	377 632 454 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
25	P. Kral, T. Pavelka and C. Cerisara, Evaluation of Dialogue Act Recognition Approaches. In: MLSP"08, Cancun, Mexico, October 2008, pp. 492 - 497, ISSN : 1551-2541, ISBN : 978-1-4244-2375-0.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/26/2008**

Název výsledku

jSynt: A Czech Text-to-Speech System written in JAVA

Abstrakt

This paper deals with a speech synthesis. Speech synthesis is an artificial production of speech by a computer. Speech synthesis from a text is called Text-to-Speech (TTS) synthesis. The main goal of this paper is to propose and implement a TTS system based on the MBROLA (Multiband Resynthesis Overlap-Add) project. We implement jSynt tool, a TTS system written in java. The first version includes two main languages, Czech and English. However the design of the system is general enough to fit other languages. The resulting synthesized speech is understandable, but not very natural.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Návrh a implementace jSynt TTS systému založeném na systému MBROLA.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Syntetizovaná řeč je srozumitelná, není ale příliš přirozená.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Král Pavel Ing. PhD.**
Spojení	377 632 454 pkral@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň liks.fav.zcu.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
26	P. Kral and K. Ekstein, jSynt: A Czech Text-to-Speech System written in JAVA, in 9th International PhD Workshop on Systems and Control, Izola, Slovenia, October 2008, ISBN : 978-961-264-003-3.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/27/2008**

Název výsledku

Návrh dialogového systému pro speciální použití

Abstrakt

Der Fahrer kommuniziert heute mit einem Navigationssystem meistens durch den Kontaktbildschirm und durch diesen Kontakt gibt er dem System Befehle zur Ausführung von bestimmten Aktionen. Wenn wir diesen Kontakt durch eine einfache, aber natürliche Dialogführung mit dem System ersetzen wollen, ist es notwendig, die klassische Benutzerschnittstelle durch ein benutzerfreundliches Dialogsystem auszutauschen und für dieses Dialogsystem eine einfache Dialogführung zu entwickeln. Beim Entwurf eines solches Systems müssen die allgemeinen Prinzipien der Dialogführung respektiert werden, aber unter Berücksichtigung jener Bedingungen, die der fahrende Wagen dem Fahrer stellt und die mit heutigen technischen Mittel realisierbar sind. Ziel des Entwurfs des erfolgreichen Dialogsystems (im allgemeinen) ist die Entwicklung der benutzerfreundlichen Schnittstelle.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Návrh a pokusná realizace automobilového navigačního systému, se kterým řidič komunikuje omezenou množinou přirozeného jazyka. Cílem aplikace bylo osvobození řidiče od nutnosti pohlížet na obrazovku navigačního systému a zvýšit tak bezpečnost silničního provozu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Konkrétní realizace vysoce specializovaného dialogového systému, který bude široce prakticky využit. Využití systému se předpokládá ve vozech fy Škoda a VW.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Matoušek Václav Prof. Ing. CSc.**
Spojení	377632471 377632402 matousek@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
27	Matoušek, V., Nestorovič, T.: Entwurf der Sprachkommunikation mit einem Auto-Navigationssystem und ihre Implementation in der VoiceXML Sprache. In: Proceedings of the Int. Conference DAGA 2008, Dresden, TUD Verlag, Dresden, März 2008	D – článek ve sborníku (RIV 2009)	NEM

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/28/2008**

Název výsledku

An automatic user-independent real-time facial expression recognition system – ARFE

Abstrakt

Automatický uživatelský nezávislý systém rozpoznávání výrazu obličeje - ARFE. Velkou výhodou a v podstatě nejzajímavější vlastností tohoto systému v porovnání s existujícími systém je plná automatizace, která je nezbytnou podmínkou pro komunikační rozhraní člověk - stroj. Je to multiuživatelský systém pracující ve dvou módech: statickém - snímky a dynamickém - snímky a video.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- AD, 2.- JD, 3.- JC, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Velkou výhodou a v podstatě nejzajímavější vlastností tohoto systému v porovnání s existujícími systém je plná automatizace, která je nezbytnou podmínkou pro komunikační rozhraní člověk - stroj.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Automatický uživatelský nezávislý systém rozpoznávání výrazu obličeje, autorizovaný software

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Klečková Jana Doc. Dr.Ing.**
Spojení	377632421 377632402 kleckova@kiv.zcu.cz
Organizace	49777513 Západočeská univerzita v Plzni Univerzitní 8 30614 Plzeň www.kiv.zcu.cz/~kleckova

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
28	Dokumentace je dostupná na http://www.kiv.zcu.cz/vyzkum/software/	R – software (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/29/2008**

Název výsledku

A Distributed Database System for Developing Ontological and Lexical Resources in Harmony.

Abstrakt

V článku je popsáno vytváření nové lexikální databáze holandštiny, projekt Cornetto, která je propojena s anglickými synsety a formální ontologií. Databáze Cornetto je založena na dvou existujících elektronických slovnících Referentie Bestand Nederlands (RBN) a holandský wordnet (DWN). V RBN jsou obsaženy také informace typu FrameNet pro holandštinu a DWN je strukturován jako anglický wordnet. V databázi Cornetto existují tři různé kolekce záznamů pro lexikální jednotky, synsety a termíny ontologie. Všechny tři jsou propojeny a záměrem projektu je úprava vztahů mezi nimi. Je představena také organizace a pracovní postupy projektu. Dále je popsán návrh a implementace nového nástroje pro lexikografickou práci. Nástroje jsou založeny na platformě DEB jako speciální klienty pro editor wordnetů DEBVisdic.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Úprava vztahů mezi kolekcemi pro lexikální jednotky, synsety a termíny ontologie, představení organizace a pracovních postupů projektu a návrh a implementace nového nástroje pro lexikografickou práci.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Návrh a implementace nového nástroje pro lexikografickou práci.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. PhD.**
Spojení	+420 549 49 4377 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/30/2008**

Název výsledku

Computer Processing of Czech Syntax and Semantics

Abstrakt

Tato práce prezentuje výsledky dosažené při řešení vybraných výzkumných projektů zpracování přirozeného jazyka v Centru ZPJ na Fakultě informatiky Masarykovy univerzity. Hlavním tématem těchto projektů je syntaktická a sémantická analýza vět přirozeného jazyka se zaměřením na češtinu. Vedoucím uvedených projektů je Aleš Horák. Po úvodu je ve druhé kapitole popsána dosavadní tříletá práce na velkém lexikonu českých slovesných valencí VerbaLex, kde valence jsou uloženy ve formě tzv. komplexních valenčních rámců. Po této části následuje detailní popis vyvinutých nástrojů pro práci s tímto i jinými jazykovými zdroji. Jedná se o nástroje VisDic, DEBVisDic, DEBDict, PRALED a další. Tyto nástroje byly (a stále jsou) používány v jazykových výzkumných projektech po celém světě. Následující kapitola ukazuje poslední vývoj syntaktického analyzátoru synt, který je jedním z dlouhodobých projektů Centra ZPJ. Kromě detailního popisu vnitřních technik a formátů systému synt uvádíme také srovnání s několika dalšími syntaktickými analyzátory přirozeného jazyka, kde ukazujeme, že synt je minimálně srovnatelný s nejlepšími současnými analyzátory. Ve čtvrté kapitole shrnujeme pokrok ve vývoji Algoritmu normální translace (NTA) pro transparentní intenzionální logiku (TIL). Popisujeme zde metody a techniky pro automatický překlad věty přirozeného jazyka na její význam ve formě konstrukce transparentní intenzionální logiky. Uvedený popis není ještě kompletní, ale zaměřili jsme se na vybrané problematické jevy, u kterých uvádíme příklady řešení nebo dokonce prototypové implementace. Poslední kapitola tohoto textu poskytuje detaily projektu zaměřeného na inteligentní metody pro zvýšení spolehlivosti elektrických sítí. Tento projekt zahrnuje jako jednu ze svých částí vývoj komunikačního rozhraní člověk-stroj pro dialogy ze specifické znalostní domény elektrorozvodných systémů (ERS).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Nejnovější informace o postupu prací na výzkumných projektech centra ZPJ FI MU.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přehled výsledků dosažných v rámci řešení výzkumných projektů zpracování přirozeného jazyka v Centru ZPJ na Fakultě informatiky Masarykovy univerzity.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. PhD.**
Spojení	549 49 4377 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 B www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
30	Horák, Aleš. Computer Processing of Czech Syntax and Semantics. 1st edition. Brno, Czech Republic : Librix.eu, 2008. 241 s. 1st edition. ISBN 978-80-7399-375-7	B – odborná kniha (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/31/2008**

Název výsledku

Can Complex Valency Frames be Universal?

Abstrakt

Článek se věnuje komplexním valenčním rámcům v databázi VerbaLex a dospívá k výsledku, ze tyto rámce mají univerzální povahu pro specifikaci argumentově predikátové struktury nejen pro češtinu, ale i pro další jazyky jako bulharština, rumunština nebo angličtina. Výsledek má význam pro multilinguální přístup k webu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Výsledek, že komplexní valenční rámce mají univerzální povahu pro specifikaci argumentově predikátové struktury nejen pro češtinu, ale i pro další jazyky jako bulharština, rumunština nebo angličtina.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Použitelnost komplexních valenčních rámců pro specifikaci argumentově predikátové struktury nejen pro češtinu, ale i pro další jazyky.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pala Karel doc. PhDr. CSc.**
Spojení	+420 549 49 5616 pala@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
31	Horák, Aleš, Rambousek, Adam, Maks, Isa, Segers, Roxane, Vossen, Piek, van der Vliet, Hennie. Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology. In The 18th International Congress of Linguists. Seoul, Republic of Korea : Korea University, 2008. od s. 190-191, 2 s.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/32/2008**

Název výsledku

Cornetto Tools and Methodology for Interlinking Lexical Units, Synsets and Ontology

Abstrakt

Projekt Cornetto vytváří lexikálně sémantickou databázi holandštiny. Databáze spojuje Wordnet s informacemi v podobě FrameNet. Data jsou odvozena z dvou existujících zdrojů: Dutch Wordnet a Referentie Bestand Nederlands. Pro uložení a editaci této kompletní databáze je použita platforma Dictionary Editor and Browser. Vazby na WordNet umožňují uplatnit multilinguální přístupy k webu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Projekt Cornetto vytváří lexikálně sémantickou databázi holandštiny. Databáze spojuje Wordnet s informacemi v podobě FrameNet. Data jsou odvozena z dvou existujících zdrojů: Dutch Wordnet a Referentie Bestand Nederlands.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Lexikálně sémantická databáze holandštiny projektu Cornetto.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. PhD.**
Spojení	+420 549 49 4377 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
32	Pala, Karel, Svoboda, Lukáš, Šmerk, Pavel. Czech MWE Database. In Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 08). Marrakech, Morocco : European Language Resources Association (ELRA), 2008. s. 1-5. ISBN 2-9517408-4-0.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/33/2008**

Název výsledku

Czech MWE Database

Abstrakt

Článek popisuje strukturu a obsah české databáze víceslovných výrazů obsahující v současnosti více než 160 000 položek a porovnává ji s daty Českého národního korpusu. Dále je navrženo, jak databázi doplňovat pomocí Word Sketch Engine.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Způsob dopňování databáze víceslovných výrazů pomocí Word Sketch Engine a její porovnání s daty Českého národního korpusu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Porovnání české MWE databáze s daty Českého národního korpusu a návrh na její doplňování pomocí Word Sketch Engine.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Pala Karel doc. PhDr. CSc.**
Spojení	+420 549 49 5616 pala@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
33	Pomikálek, Jan - Rychlý, Pavel. Detecting Co-Derivative Documents in Large Text Collections. In Proceedings of the Sixth International Language Resources and Evaluation (LREC"08). Marrakech, Morocco : European Language Resources Association (ELRA), 2008. od s. 132-135, 3 s. ISBN 2-9517408-4-0.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/34/2008**

Název výsledku

Detecting Co-Derivative Documents in Large Text Collections.

Abstrakt

Analyzovali jsme algoritmus SPEX (Bernstein a Zobel, 2004) pro detekci blízkých dokumentů s použitím duplicitních n-gramů. Přestože zcela souhlasíme s tvrzením, že zanedbání unikátních n-gramů může vést ke značenému zvýšení efektivity a škálovatelnosti procesu detekce blízkých dokumentů, objevili jsme závažné nedostatky ve způsobu, kterým SPEX vyhledává duplicitní n-gramy. Paměťové nároky na výpočet blízkých dokumentů mohou být sníženy až na 1%, použijeme-li pouze duplicitní n-gramy, avšak SPEX potřebuje přibližně 40x více paměti pro výpočet samotného seznamu duplicitních n-gramů. Celkové paměťové nároky tedy nejsou dostatečně nízké na to, aby byl algoritmus prakticky použitelný pro velmi velké kolekce. Navrhli jsme řešení tohoto problému s použitím externího řazení s řazením v paměti pomocí sufixového pole a komprese dočasných souborů. Navržený algoritmus pro výpočet duplicitních n-gramů vyžaduje pevné množství paměti pro vstup libovolné velikosti.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Vylepšení algoritmu SPEX pro detekci blízkých dokumentů s použitím duplicitních n-gramů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Snížení paměťových nároků algoritmu pro detekci blízkých dokumentů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Rychlý Pavel Mgr. Ph.D.**
Spojení	+420 549 49 6399 pary@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
34	Kopeček, Ivan - Ošlejšek, Radek. Dialogue-Based Processing of Graphics and Graphical Ontologies. 11th International Conference, TSD 2008. Berlin : Springer-Verlag, Brno 2008.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/35/2008**

Název výsledku

Dialogue Based Text Editing

Abstrakt

Článek popisuje základní principy editace textu pomocí dialogu. Nejdříve je představen algoritmus pro dělení textu včetně jeho rozšíření. Dále je ukázáno dialogové rozhraní pro zpracování textu, které spolupracuje se syntetizérem řeči. Byly navrženy základní funkce a také formulovány nejdůležitější problémy a jejich možná řešení.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Rozšíření algoritmu pro dělení textu, dialogové rozhraní pro editaci spolupracující se syntetizérem řeči a formulace a možná řešení nejdůležitějších problémů tohoto procesu.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Dialogové rozhraní pro editaci textu a možná řešení některých problemových dílčích úkolů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Plhák Jaromír Mgr.**

Spojení xplhak@fi.muni.cz

Organizace 00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
35	Ivanova, Kremena - Heid, Ulrich - Schulte im Walde, Sabine - Kilgarriff, Adam - Pomikálek, Jan. Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In Proceedings of the Sixth International Language Resources and Evaluation (LREC''08). Marrakech, Morocco : European Language Resources Association (ELRA), 2008. od s. ?, 7 s. ISBN 2-9517408-4-0.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/36/2008**

Název výsledku

Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case.

Abstrakt

Word sketches jsou součástí korpusového manažeru Sketch Engine. Reprezentují shrnutí gramatického a kolokačního chování slov, automaticky odvozené z korpusu. Pro vytvoření word sketches je kromě korpusu zapotřebí rovněž tzv. sketch grammar, mělká gramatika založená na regulárních výrazech nad morfologickými značkami. Tento článek představuje sketch grammar pro Němčinu, jazyk s poměrně volným slovosledem, který vykazuje značné známky sykretismu, a vyhodnocuje její úspěšnost, což dosud nebylo provedeno pro žádnou jinou sketch grammar. Vyhodnocení se zaměřuje na jmenné fráze jakožto zásadní část německé gramatiky. Představujeme různé verze definic jmenných frází a jejich vliv na přesnost a úplnost výsledků.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Tento článek představuje sketch grammar pro Němčinu, jazyk s poměrně volným slovosledem, který vykazuje značné známky sykretismu, a vyhodnocuje její úspěšnost, což dosud nebylo provedeno pro žádnou jinou sketch grammar. Vyhodnocení se zaměřuje na jmenné fráze jakožto zásadní část německé gramatiky. Představujeme různé verze definic jmenných frází a jejich vliv na přesnost a úplnost výsledků.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přínosem výsledku je sketch grammar pro Němčinu, jazyk s poměrně volným slovosledem, který vykazuje značné známky sykretismu, a vyhodnocuje její úspěšnost, což dosud nebylo provedeno pro žádnou jinou sketch grammar.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Pomikálek Jan RNDr.**

Spojení xpomikal@fi.muni.cz

Organizace 00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
36	Jakubíček, Miloš. Extraction of Syntactic Structures Based on the Czech Parser Synt. In Proceedings of Recent Advances in Slavonic Natural Language Processing 2008. Brno : Masaryk University, 2008. s. 56-62. ISBN 978-80-210-4741-9.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/37/2008**

Název výsledku

Extraction of Syntactic Structures Based on the Czech Parser Synt

Abstrakt

Článek popisuje využití syntaktického analyzátoru Synt k získání informací o syntaktických strukturách z běžného českého textu. Tyto struktury z pohledu analýzy zpravidla odpovídají neterminálům v gramatice využívané parserem k nalezení platných odvození zadané věty. Tento parser byl rozšířen tak, aby nabízel několik způsobů, jak využít jeho masivně víceznačný výstup k jednoznačné extrakci syntaktických struktur. Za tímto účelem byly zapojeny i některé dosud nevyužité výsledky syntaktické analýzy vedoucí ke zpřesnění morfologické analýzy a tím i k většímu rozlišení různých syntaktických (pod)struktur. Závěrem je představeno využití pro hrubou extrakci slovesných valencí.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Článek popisuje využití syntaktického analyzátoru Synt k získání informací o syntaktických strukturách z běžného českého textu.Tento parser byl rozšířen tak, aby nabízel několik způsobů, jak využít jeho masivně víceznačný výstup k jednoznačné extrakci syntaktických struktur. Za tímto účelem byly zapojeny i některé dosud nevyužité výsledky syntaktické analýzy vedoucí ke zpřesnění morfologické analýzy a tím i k většímu rozlišení různých syntaktických (pod)struktur.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Rozšíření syntaktického analyzátoru Synt k získání informací o syntaktickýc strukturách z běžného českého textu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Jakubíček Miloš Bc.**
Spojení	mjakubicek@mail.muni.cz
Organizace	00216224 Masarykova univezita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
37	Rychlý, Pavel - Husák, Miloš - Kilgarriff, Adam - Rundell, Michael - McAdam, Katy. GDEX: Automatically finding good dictionary examples in a corpus. In Proceedings of the XIII EURALEX International Congress. 1. vyd. Barcelona : Institut Universitari de Lingüística Aplicada, 2008. od s. 425-432, 7 s. ISBN 9788496742673.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/38/2008**

Název výsledku

GDEX: Automatically finding good dictionary examples in a corpus.

Abstrakt

Příklady užití slov či frází v heslem slovníků jsou cenným zdrojem informací. Vytváření vhodných příkladů je ovšem náročná manuální práce. Článek popisuje algoritmus automatické identifikace vět z korpusu, které mohou sloužit jako dobré příklady užití pro slovníková hesla.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Algoritmus automatické identifikace vět z korpusu, které mohou sloužit jako dobré příklady užití pro slovníková hesla.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Algoritmus automatické identifikace vět z korpusu, které mohou sloužit jako dobré příklady užití pro slovníková hesla.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Rychlý Pavel Mgr. Ph.D.**

Spojení +420 549 49 6399 pary@fi.muni.cz

Organizace 00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
38	Němčík, Václav - Hlaváčková, Dana - Horák, Aleš - Pala, Karel - Úradník, Michal. Processing Czech Verbal Synsets with Relations to English WordNet. In RASLAN 2008. 2. vyd. Brno : Masarykova Univerzita, 2008. od s. 49-55, 7 s. ISBN 978-80-210-4741-9.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/39/2008**

Název výsledku

Processing Czech Verbal Synsets with Relations to English WordNet

Abstrakt

Tento článek popisuje aktuální výsledky dokončování druhé stabilní verze velkého slovníku valencí českých sloves VerbaLex. VerbaLex je vyvíjen na Centru zpracování přirozeného jazyka během posledních tří let. Cílem nynější fáze vývoje tohoto unikátního zdroje jazkových dat je jeho úplné propojení s nejvýznamnější světovou sémantickou sítí, Princetonským WordNetem. Tento článek popisuje metodiku, jak dosáhnout tohoto cíle poloautomaticky. Dále se článek věnuje vybraným etapám přípravy tištěného slovníku obsahujícího podstatnou část VerbaLexu, s valenčními rámci uzpůsobenými do podoby přehlednější pro lidského čtenáře.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Tento článek popisuje metodiku, jak poloautomaticky dosáhnout propojení slovníku valencí VerbaLex s WordNetem. Dále se článek věnuje vybraným etapám přípravy tištěného slovníku obsahujícího podstatnou část VerbaLexu, s valenčními rámci uzpůsobenými do podoby přehlednější pro lidského čtenáře.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Tento článek popisuje metodiku, jak poloautomaticky dosáhnout propojení slovníku valencí VerbaLex s WordNetem. Dále se článek věnuje vybraným etapám přípravy tištěného slovníku obsahujícího podstatnou část VerbaLexu, s valenčními rámci uzpůsobenými do podoby přehlednější pro lidského čtenáře.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Němčík Václav Mgr.**

Spojení xnemcik@fi.muni.cz

Organizace 00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
39	Sojka, Petr - Horák, Aleš. Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008. Edited by Sojka P., Horák A. Vyd. prvni. Brno : Masaryk University, 2008. 110 s. RASLAN Proceedings. ISBN 978-80-210-4741-9.	B – odborná kniha (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/40/2008**

Název výsledku

Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008.

Abstrakt

borník obsahuje 14 odborných příspěvků z oblastí jazykové morfologie, syntaktické a sémantické analýzy, softwarových nástrojů pro zpracování textu a lexikální sémantiky. Je určen všem vědcům a studentům zajímajícím se o komputační lingvistiku v oblasti jazykového inženýrství slovanských jazyků.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

14 původních článků z oblastí morfologie, syntaktické a sémantické analýzy a softwareových nástrojů pro zpracování textu a lexikální sémantiky.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Přehled nových poznatků z oblasti morfologie, syntaktické a sémantické analýzy a softwareových nástrojů pro zpracování textu a lexikální sémantiky.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr doc. RNDr. Ph.D.**
Spojení	+420 549 49 6966 sojka@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
40	Němčík, Václav - Pala, Karel - Hlaváčková, Dana. Semi-automatic Linking of New Czech Synsets Using Princeton WordNet. In Intelligent Information Systems XVI, Proceedings of the International IIS""08 Conference. Warszawa : Academic Publishing House EXIT, 2008. od s. 369-374, 6 s. ISBN 978-83-60434-44-4.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/41/2008**

Název výsledku

Semi-automatic Linking of New Czech Synsets Using Princeton WordNet

Abstrakt

Tento příspěvek se týká rozšiřování českého WordNetu o synsety obsažené v databázi českých valenčních rámců Verbalex. V této souvislosti je jedním z hlavních cílů zavěsit nově přidávané synsety k jejich hyperonymům a přiřadit k nim odpovídající synsety v anglickém Princeton WordNetu. Abychom ušetřili lexikografům rutinní úkony a zefektivnili jejich práci, vyvinuli jsme WordNet Asistenta, softwarový nástroj, který pomáhá nalézt relevantní synset(y) v již existujících datových strukturách. Tento nástroj představuje dle našich zkušeností významné usnadnění práce při začleňování nových synsetů do českého WordNetu a domníváme se, že bude vítanou pomocí při rozšiřování WordNetů pro další jazyky.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

WordNet Asistenta, softwarový nástroj, který pomáhá nalézt relevantní synset(y) v již existujících datových strukturách.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

WordNet Asistent, softwarový nástroj, který pomáhá nalézt relevantní synset(y) v již existujících datových strukturách při začleňování nových synsetů do českého WordNetu.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Němčík Václav Mgr.**

Spojení xnemcik@fi.muni.cz

Organizace 00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
41	Kovář, Vojtěch - Jakubíček, Miloš. Test Suite for the Czech Parser Synt. In Proceedings of Recent Advances in Slavonic Natural Language Processing 2008. Brno : Masaryk University, 2008. s. 63-70. ISBN 978-80-210-4741-9.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/42/2008**

Název výsledku

Test Suite for the Czech Parser Synt

Abstrakt

Článek popisuje sadu nástrojů určených pro testování syntaktického analyzátoru češtiny Synt. Zabývá se též použitými syntaktickými daty, nově vytvořeným brněnským korpusem složkových stromů.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Sada nástrojů pro testování syntaktického analyzátoru Synt a nově vytvořený korpus složkových stromů.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Sada nástrojů pro testování syntaktického analyzátoru Synt a nově vytvořený korpus složkových stromů.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Kovář Vojtěch Bc.**

Spojení xkovar3@fi.muni.cz

Organizace 00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
42	Horák, Aleš - Vossen, Piek - Rambousek, Adam. The Development of a Complex-Structured Lexicon based on WordNet. In Proceedings of the Fourth Global WordNet Conference. Szeged : University of Szeged, 2008. od s. 200-208, 9 s. ISBN 978-963-482-854-9.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/43/2008**

Název výsledku

The Development of a Complex-Structured Lexicon based on WordNet.

Abstrakt

V projektu Cornetto je vyvíjen nový komplexně strukturovaný lexikon holandštiny. Lexikon spojuje informace ze dvou elektronických slovníků - Referentie Bestand Nederlands (RBN) a Dutch WordNet (DWN). Lexikon Cornetto bude navázán na synsety anglického WordNetu a bude obsahovat podrobný popis lexikálních jednotek (morfologické, syntaktické a sémantické informace). Databáze je rozdělena do čtyř částí - lexikální jednotky, synsety, ontologie a Cornetto identifikátory. Cornetto identifikátory slouží k propojení synsetů a lexikálních jednotek, nejprve jsou vytvořeny automaticky, později manuálně upravovány. Pro práci se slovníky bylo vytvořeno speciální uživatelské rozhraní. Článek popisuje implementaci nástrojů, založených na editoru DEBVisDic. Vývoj software pro Cornetto je společným projektem Masarykovy Univerzity a University of Amsterdam.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Implementaci nástrojů, založených na editoru DEBVisDic pro projekt Cornetto.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Implementaci nástrojů, založených na editoru DEBVisDic pro projekt Cornetto.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 49 4377 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
43	Horák, Aleš - Pala, Karel - Rambousek, Adam. The Global WordNet Grid Software Design. In Proceedings of the Fourth Global WordNet Conference. Szeged : University of Szeged, 2008. od s. 194-199, 6 s. ISBN 978-963-482-854-9.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/44/2008**

Název výsledku

The Global WordNet Grid Software Design

Abstrakt

Článek představuje software pro Global WordNet Grid. Cílem Gridu je vytvořit síť volně dostupných WordNetů, propojených mezijazykovými indexy. NLP Centrum FI MU připravuje software pro vytvoření Gridu. Všechny WordNety budou uloženy na speciálně vytvořeném DEB serveru. V článku je popsána aplikace DEBGrid a různé možnosti uložení dat a řízení uživatelského přístupu k WordNetům.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Vytvoření sítě volně dostupných WordNetů, propojených mezijazykovými indexy. NLP Centrum FI MU připravuje software pro vytvoření Gridu

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Síť volně dostupných WordNetů, propojených mezijazykovými indexy

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 49 4377 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
44	Němčík, Václav. The Saara Framework: Work in Progress. In: RASLAN 2008. 2. vyd. Brno : Masarykova univerzita, 2008. od s. 11-16, 6 s. ISBN 978-80-210-4741-9.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/45/2008**

Název výsledku

The Saara Framework: Work in Progress.

Abstrakt

Určení referencí výrazů a referenčních vztahů v diskursu je jedním z nejdůležitějších úkolů, které je třeba řešit při automatickém porozumění textu. Prvním krokem k tomuto cíli je určit koreferenční třídy nad množinou referenčních výrazů. Tento článek představuje modulární systém pro automatickou analýzu anafor, který umožňuje používat různé algoritmy pro analýzu anafor a aplikovat je obecně vzato na libovolný jazyk. Funkcionalita systému je ilustrována na vybraných algoritmech založených na modelování aktivovanosti a upravených pro češtinu.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Modulární systém pro automatickou analýzu anafor, který umožňuje používat různé algoritmy pro analýzu anafor a aplikovat je obecně vzato na libovolný jazyk.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Modulární systém pro automatickou analýzu anafor.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Němčík Václav Mgr.**
Spojení	xnemcik@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
45	Horák, Aleš - Pala, Karel - Rambousek, Adam. Tools for Managing Multilingual Lexical Resources. In Procedings of the 16th International Conference Intelligent Information Systems. Zakopane, Poland : Polish Academy of Sciences, 2008. od s. 451-460, 10 s. ISBN 978-83-60434-44-4.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/46/2008**

Název výsledku

Tools for Managing Multilingual Lexical Resources

Abstrakt

Článek popisuje výsledky vývoje nástrojů pro správu vícejazyčných lexikálních zdrojů, např. jednojazyčných i vícejazyčných slovníků, terminologických slovníků, komplexních lexikografických databází nebo sémantických sítí WordNet. Všechny prezentované nástroje jsou založeny na platforme DEB, která využívá standardní XML formáty. Usilujeme také o standardizaci lexikálních zdrojů a jejich propojení. Představené nástroje jsou volně dostupné. Podrobněji jsou představeny aplikace DEBDict (obecný prohlížeč slovníků), DEBTerm (vícejazyčný terminologický slovník), Praled (česká lexikální databáze) a Visual Browser (grafický prohlížeč sémantických sítí).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Výsledky vývoje nástrojů pro správu vícejazyčných lexikálních zdrojů, např. jednojazyčných i vícejazyčných slovníků, terminologických slovníků a nebo sémantických sítí WordNet.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Aplikace DEBDict (obecný prohlížeč slovníků), DEBTerm (vícejazyčný terminologický slovník), Praled (česká lexikální databáze) a Visual Browser (grafický prohlížeč sémantických sítí).

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Horák Aleš RNDr. Ph.D.**
Spojení	+420 549 49 4377 hales@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
46	Šmerk, Pavel. Towards Czech Morphological Guesser. In Sojka, Petr - Horák, Aleš. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008. Brno : Masarykova univerzita, 2008. od s. 1-4, 4 s. ISBN 978-80-210-4741-9	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/47/2008**

Název výsledku

Towards Czech Morphological Guesser.

Abstrakt

Článek prezentuje morfologický guesser pro češtinu, který je založený na datech českého morfologického analyzátoru ajka. Konstrukce je založena na předpokladu, že nová (a tedy analyzátoru neznámá) slova se v jazyce chovají pravidelně, a že navíc tato pravidelnost může být extrahována z existujících dat. Článek popisuje jak tvorbu dat, tak fungování samotného guesseru.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Morfologický guesser pro češtinu, který je založený na datech českého morfologického analyzátoru ajka. Konstrukce je založena na předpokladu, že nová (a tedy analyzátoru neznámá) slova se v jazyce chovají pravidelně, a že navíc tato pravidelnost může být extrahována z existujících dat.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Morfologický guesser pro češtinu, který je založený na datech českého morfologického analyzátoru ajka.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Šmerk Pavel RNDr.**
Spojení	+420 549 49 4347 xsmerk@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
47	Sojka, Petr. Towards Natural Natural Language Processing. In RASLAN 2008 Proceedings. Vyd. první. Brno : Masaryk University, 2008. ISBN 978-80-210-4741-9, s. 98-100. 5.12.2008, Karlova Studánka.	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/48/2008**

Název výsledku

Towards Natural Natural Language Processing.

Abstrakt

Esej o tom jak zpracovávat přirozeně přirozený jazyk, inspirujíce se z domnělého způsobu zpracování jazyka v našich hlavách (zpracování víceznačností ap.).

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Domnělý způsob zpracování jazyka v našich hlavách.

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Domnělý způsob zpracování jazyka v našich hlavách.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Sojka Petr doc. RNDr. Ph.D.**
Spojení	+420 549 49 6966 sojka@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
48	Bártek, L., Plhák, J. - Visually Impaired Users Create Web Pages, in In Computers Helping People with Special Needs: 11th International Conference, ICCHP 2008. Berlin : Springer-Verlag, 2008. od s. 466-473, ISBN 3-540-70539-2	D – článek ve sborníku (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/49/2008**

Název výsledku

Visually Impaired Users Create Web Pages

Abstrakt

WebGen má umožnit zrakově postiženým uživatelům jednoduše a přirozeně vytvářet webové prezentace pomocí dialogu. Tento článek popisuje základní metody a principy systému WebGen, jako jsou jeho struktura a dialogové rozhraní. Článek dále obsahuje ukázku dialogu a jím vygenerované webové stránky.

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- IN, 2.- JC, 3.- JD, 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Dialogvá tvorba webovských prezentací

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Nástroj WebGen, který umožňuje dialogovou formou vytvářet některé typy webovských stránek a provazovat je do prezentací.

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno	**Bártek Luděk Mgr. Ph.D.**
Spojení	+420 549 49 3215 bar@fi.muni.cz
Organizace	00216224 Masarykova univerzita Žerotínovo náměstí 617 9 60177 Brno www.muni.cz

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk
49	Horák, Aleš, Kopeček, Ivan,Pala, Karel, Sojka, Petr: Text, Speech and Dialogue: Proceedings of the 11th International Conference TSD 2008, Brno, Czech Republic, September 8-11, 2008. Edited by Sojka P., Horák A., Kopeček I., Pala K. Berlin Heidelberg : Springer Verlag, 2008. 667 s. Lecture Notes in Computer Science Volume 5246. ISBN 978-3-540-87390-7.	B – odborná kniha (RIV 2009)	ANG

1. ZÁKLADNÍ ÚDAJE

Číslo výsledku: **2C06009/50/2008**

Název výsledku

Text, Speech and Dialogue: Proceedings of the 11th International Conference TSD 2008, Brno,

Abstrakt

Kniha obsahuje články a postery prezentované na konferenci TSD 2008

Hlavní (1) a další (2-5) obory řešení výsledku (dle číselníku CEP, RIV)

1.- , 2.- , 3.- , 4.- , 5.-

2. INOVAČNÍ ASPEKTY

Popis inovačních aspektů daného výsledku

Publikace o vědecké konferenci přinášející aktuální výsledky výzkumu v oblasti NLP

3. PŘÍNOSY

Popis konkrétních přínosů daného výsledku pro jeho uživatele

Kvalitní a aktuální informace o současném výzkumu na poli NLP

4. KONTAKTNÍ ÚDAJE GARANTA VÝSLEDKU

Celé jméno **Sojka Petr doc. RNDr. PhD.**

Spojení sojka@fi.muni.cz

Organizace

5. DOSTUPNÁ DOKUMENTACE

Číslo	Název dokumentu	Typ	Jazyk

4.1.3. PLNĚNÍ DÍLČÍCH CÍLŮ

4.1.3.1. ZPRÁVA O DOSAŽENÍ DÍLČÍHO CÍLE

Číslo dílčího cíle	2
Název dílčího cíle	Návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka.
Plánované datum dosažení dílčího cíle	31.12.2008

INDIKÁTORY DOSAŽENÍ VÝSTUPU - SKUTEČNĚ DOSAŽENÉ

V průběhu roku 2008 bylo dokončeno zpracování aktivity č.2 - návrh formalismů a modelů komunikace s www prostředím na bázi přirozeného jazyka. Bylo dosaženo následujících významných výsledků:
- nově navržená metoda sdružující citační analýzu se sítí spoluautorství,
- nové algebraické metody pro redukci informací,
- nová metoda pro automatickou detekci plagiátů ve vícejazyčném prostředí,
- optimalizovaná metoda umožňující extrakci N-gramů z rozsáhlých textových dat,
- dvě nové metody pro extrakci užitečné informace z Webu,
- nová metoda sémantické analýzy lexikálních tříd,
- metoda kategorizace česky psaných dokumentů metodou WEBSOM,
- zpracování a kategorizace česky psaných dokumentů umělou neuronovou sítí,
- nově vyvinutá hybridní metoda automatického rozpoznávání řeči,
- nová metoda vyhodnocování kvality automatického rozpoznávání dialogových aktů,

PROSTŘEDKY OVĚŘENÍ VÝSTUPU - SKUTEČNĚ DOSAŽENÉ

Výše uvedené nově navržené metody byly ověřovány prostřednictvím následujících ověřovacích metod a softwaru:
- nový způsob využití umělých neuronových sítí pro modelování kontextově závislých jednotek jazyka,
- nová verze hybridního rozpoznávače řeči založená na použití speciálního typu umělé neuronové sítě,
- software pro automatizovaný, scénářem řízený sběr korpusového materiálu,
- rozšíření možností počítačového zpracování sémantiky přirozeného jazyka z pohledu neurověd, lingvistiky a informatiky,
- implementace N-best dekódéru pro rozpoznávač JLASER,
- návrh a rozsáhlé testování automatického rozpoznávače řeči pomocí kontextově závislých slabik,
- návrh a implementace TTS systému jSynt založeném na systému MBROLA.

4.1.4. REDAKČNĚ UPRAVENÁ ZPRÁVA

Projekt má za cíl vývoj nástrojů pro komunikaci s Webem v přirozeném jazyce. V r. 2008 byla dokončena většina korpusů

pro trénování a verifikaci vlastností navrženého systému,

dokončena byla nová verze rozpoznávače, implementována byla řada různých algoritmů pro zpracování obsahu Webu prostřednictvím přirozeného jazyka a nadále budou testovány jejich vlastnosti.

K významným výsledkům dosaženým v průběhu roku 2008 patří zejména:

- nová metoda pro vyhodnocování autoritativnosti výzkumníků a výzkumných skupin,
- nově navržená metoda sdružující citační analýzu se sítí spoluautorství,
- nové algebraické metody pro redukci informací,
- nová metoda pro automatickou detekci plagiátů ve vícejazyčném prostředí,
- optimalizovaná metoda umožňující extrakci N-gramů z rozsáhlých textových dat,
- dvě nové metody pro extrakci užitečné informace z Webu,
- nová metoda sémantické analýzy lexikálních tříd,
- metoda kategorizace česky psaných dokumentů metodou WEBSOM,
- zpracování a kategorizace česky psaných dokumentů umělou neuronovou sítí,
- rozšíření možností počítačového zpracování sémantiky přirozeného jazyka z pohledu neurověd, lingvistiky a informatiky,
- nový způsob využití umělých neuronových sítí pro modelování kontextově závislých jednotek jazyka,
- nová verze hybridního rozpoznávače řeči založená na použití speciálního typu umělé neuronové sítě,
- software pro automatizovaný, scénářem řízený sběr korpusového materiálu,
- implementace N-best dekódéru pro rozpoznávač JLASER,
- návrh a rozsáhlé testování automatického rozpoznávače řeči pomocí kontextově závislých slabik,
- nově vyvinutá hybridní metoda automatického rozpoznávání řeči,
- nová metoda vyhodnocování kvality automatického rozpoznávání dialogových aktů,
- návrh a implementace TTS systému jSynt založeném na systému MBROLA.
- formalizace lexikální databáze VerbaLex
- modelu české syntaxe na základě korpusu syntaktických stromů
- návrh formalismu pro práci s konstrukcemi TILu
- návrh a implementace morfologického guesseru
- detekce plagiátů s využitím sémantických znalostí
- algoritmy pro vytváření grafiky a webovských prezentací prostřednictvím dialogových systémů
- výsledky v oblasti klasifikace matematických textů

4.1.5. PLNĚNÍ PODMÍNEK PROGRAMU

Plnění specifických podmínek programu - se pro projekty NPV II nezpracovává. Pro projekty NPVII specifické podmínky ve vyhlášení programu nebyly formulovány.

4.1.6. PLNĚNÍ SMLOUVY O SPOLUPRÁCI

Na základě vymezených základních práv (viz uzavřená smlouva upravující vztahy mezi příjemcem a spolupříjemcem) příjemce poskytnul spolupříjemci finanční dotaci přímým převodem na stanovený účet Masarykovy univerzity, náklady na projekt byly vedeny v oddělené evidenci obou spolupracujících subjektů.
Uzavřená smlouva o spolupráci je plněna beze zbytku, plánované finanční prostředky byly vyčerpány - viz odstavec 2.3.2.

4.2. DALŠÍ PŘÍLOHY - rok 2008

4.2.1. Odborné a věcné přílohy zprávy - seznam

	Pořadí	Soubor
	1	**Seznam publikovaných prací ze ZČU v Plzni** Soubor obsahuje výčet publikovaných prací, které byly zpracovány a uveřejněny v rámci řešení projektu v roce 2008. Publikace_2C06009.pdf *(152 kB)*
	3	**Publikace Centra ZPJ k projektu** Soubor obsahuje seznam publikací k projektu za r. 2008 literNPV208fin.rtf *(20 kB)*

4.2.2. Ostatní (např. možné využití výsledků) - seznam

	Pořadí	Soubor
	1	**Soubor nejvýznamnějších publikací** V příloze Publikace.zip přikládáme dvacet významných publikací, které byly publikovány v průběhu roku 2008. Publikace.zip *(6500 kB)*

4.2.3. Zápisy z projednání (oponentní řízení, atd.) - seznam

	Pořadí	Soubor
	1	**Zápis z koordinačního semináře řešitelů projektu** Přiložený soubor obsahuje stručný zápis z koordinačního semináře řešitelů projektu, který proběhl ve dnech 31.10. - 1.11.2008 ve středisku MU Brno v Cikháji. Cílem semináře bylo vyhodnocení dosažených výsledků za první polovinu období řešení projektu a stanovení konkrétních cílů pro druhé (závěrečné) časové období - viz přiložený zápis. Zapis_Cikhaj.doc *(37 kB)*

4.2.4. Zápisy a dokumenty z jednání s administrátory programu poskytovatele - seznam

	Pořadí	Soubor
	1	**Jednání v průběhu roku 2008 neproběhla.** *(kB)*

4.2.5. Zápisy z jednání Rady projektu (Centra) - seznam

Příloha 4.2.5. Zápisy z jednání Rady projektu (Centra) - se pro tento program nezpracovává.

4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodněním - seznam

Příloha 4.2.6. Návrh dodatku ke smlouvě na řešení projektu se zdůvodnění - se pro tento program nezpracovává.